ARTIFICIAL INTELLIGENCE APPLICATIONS IN EARLY DIAGNOSIS OF SEPSIS DISEASE

SEPSİS HASTALIĞININ ERKEN TANISINDA YAPAY ZEKA UYGULAMALARI

ÖZNUR ESRA PAR

PROF. DR. EBRU SEZER

Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment to Requirements

for the Award of the Degree of Doctor of Philosophy

in Computer Engineering.

2023

ABSTRACT

ARTIFICIAL INTELLIGENCE APPLICATIONS IN EARLY DIAGNOSIS OF SEPSIS DISEASE

Öznur Esra PAR

Doctor of Philosophy, Department of Computer Engineering

Supervisor: Prof. Dr. Ebru SEZER

Co- Supervisor: Prof. Dr. Hayri SEVER

September 2023, 210 pages

Sepsis is a major cause of death in intensive care units worldwide. Early diagnosis and treatment are crucial for improving patient survival and reducing organ dysfunction. Combining sepsis research and computer science advances creates predictive models for identifying patients at risk, enabling earlier intervention and better outcomes. The connected model, proposed one was used to evaluate machine learning algorithms across patient age cohorts (infant, elder, and all age) within the context of the study. The connected model, which is thought to consider the possibility of the patient's previous condition and/or conditions, in situations like illness that spreads over time, was compared with the non-connected model, which is thought to depend only on the current situation. The connected Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), Convolution Neural Network (CNN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) machine learning models created for various patient

cohorts improved the study's ability to predict sepsis in a shorter amount of time. According to analysis of proposed model, sepsis can be predicted in the infant patient cohort at 4th hour and in the elder and all age patient cohorts at 3rd hour.

Keywords: sepsis; early prediction; machine learning, artificial intelligence, early alert.

ÖZET

SEPSİS HASTALIĞININ ERKEN TANISINDA YAPAY ZEKA UYGULAMALARI

Öznur Esra PAR

Doktora, Bilgisayar Mühendisliği Bölümü

Tez Danışmanı: Prof. Dr. Ebru SEZER

Eş Danışman: Prof. Dr. Hayri SEVER

Eylül 2023, 210 sayfa

Sepsis, tüm dünyada yoğun bakım ünitelerinde ölümün başlıca nedenlerinden biridir. Erken tanı ve tedavi, hasta sağkalımını iyileştirmek ve organ disfonksiyonunu azaltmak için hayati öneme sahiptir. Sepsis araştırmalarını, bilgisayar bilimlerindeki ilerlemelerle birleştirmek, risk altındaki hastaları belirlemek için öngörücü modeller oluşturmaktadır. Bu sayede daha erken müdahale ve daha iyi sonuçlar sağlanmaktadır. Çalışma kapsamında önerilen bağlantılı model, çalışmanın bağlamı içinde hasta yaş gruplarına (bebek, yaşlı ve her yaş) göre makine öğrenimi algoritmalarını değerlendirmek için kullanılmıştır. Zamanla bağlı değişiklik gösteren hastalık durumlarında, hastanın önceki durumlarının dikkate alınıldığı bağlantılı model, sadece mevcut duruma odaklanan bağlantısız modellerle karşılaştırılmıştır. Çeşitli hasta kohortları için oluşturulan bağlantılı Çok Katmanlı Algılayıcı (MLP), Uzun Kısa Vadeli Hafıza (LSTM), Evrişimli Sinir Ağı (CNN), Rastgele Orman (RF) ve Aşırı Gradyan Artırma (XGBoost) makine öğrenme modelleri, sepsisin daha kısa sürede tahmin edilmesi yeteneğini iyileştirmiştir. Önerilen modelin analizine göre, sepsis bebek hasta kohortunda 4. saatte ve yaşlı ile tüm yaş hasta kohortlarında 3. saatte öngörülebilmektedir.

Anahtar Kelimeler: sepsis: erken tahmin; makine öğrenimi; yapay zeka; erken uyarı.

TEŞEKKÜR

Akademik yolculuğumun her aşamasında bana rehberlik eden, bilgilerini ve deneyimlerini cömertçe paylaşan değerli danışmanım Prof. Dr. Ebru Sezer'e ve eş danışmanım Prof. Dr. Hayri Sever'e derin şükranlarımı sunarım. Bilimsel görüşleri, yapıcı eleştirileri ve değerli yönlendirmeleriyle tezimin gelişimine katkıda bulunan Prof. Dr. Hasan Oğul'a ve Doç. Dr. Murat Aydos'a teşekkür ederim. Ayrıca, çalışmamı titizlikle inceleyen ve değerli görüşleriyle zenginleştiren tez jüri üyeleri Prof. Dr. Suat Özdemir ve Doç Dr. Mehmet Serdar Güzel'e da içten teşekkürlerimi sunarım. Bilgileri, deneyimleri ve eleştirel bakış açıları, bu tezin son halini almasında önemli bir rol oynamıştır.

Ailemin benzersiz desteği ve sabrı olmasaydı bu çalışmayı tamamlamam mümkün olmazdı. Kız kardeşim Nagihan Selvi Uysallı'ya, beni her zaman desteklediği ve inandığı için; anneme, sonsuz sevgisi, sabrı ve anlayışı için içtenlikle teşekkür ederim. Aileme, akademik hedeflerimi takip etme konusundaki teşviki ve anlayışı için minnettarım.

En önemlisi, bu sürecin her anında yanımda olan, bana güç veren ve yaşamıma neşe katan oğlum Aykut Yiğit Par'a - bu tezin gerçek ilham kaynağına - en derin sevgi ve şükranlarımı sunuyorum. Onun varlığı, bu akademik serüvenimin en büyük motivasyonu olmuştur.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SYMBOLS AND ABBREVIATIONS

Symbols

Hg	Mercury
m	milli-
p(~)	Probability of Specific Measurements Triggering Sepsis Onset
PaO2	Partial Pressure of Oxygen
SpO2	Peripheral Oxygen Saturation
μ	micro-
k	kilo-

Abbreviations

AdaBoost	Adaptive Boosting
AI	Artificial Intelligence
AIDS	Acquired Immune Deficiency Syndrome
AKI	Acute Kidney Injury
ANN	Artificial Neural Network
APACHE	Acute Physiology and Chronic Health Evaluation
ARDS	Acute Respiratory Distress Syndrome
AUC	Area Under the Curve
AUPRC	Area Under the Precision-Recall Curve
AVPU	Alert, Verbal, Pain, Unresponsive
bpm	Beats per minute
bun	Blood Urea Nitrogen
CART	Classification and Regression Tree

CDSA	Clinical Decision Support Algorithms
CNN	Convolutional Neural Networks
COVID-19	Coronavirus Disease 2019
CPAP	Continuous Positive Airway Pressure
CPT	Current Procedural Terminology
DFN	Dense Fusion Network
diasbp	Diastolic Blood Pressure
DL	Deep Learning
DNN	Deep Neural Networks
DRG	Diagnosis Related Group
DSPA	Deep SOFA-Sepsis Prediction Algorithm
ECG	Electrocardiogram
ED	Emergency Department
EHR	Electronic Health Record
eICU-CRD	eICU Collaborative Research Database
EMR	Electronic Medical Record
ESICM	European Society of Intensive Care Medicine
et al.	et alii ("and others")
FCNN	Fully Connected Neural Networks
FN	False Negatives
FNR	False Negative Rate
FP	False Positives
FPR	False Positive Rate
GBDT	Gradient Boosting Decision Tree
Gboost	Gradient Boosting
GCS	Glasgow Coma Scale

GRU	Gated Recurrent Units
ICD9	International Classification of Diseases, Ninth Revision
ICU	Intensive Care Unit
ID	Identification
IHM	In-Hospital Mortality
kg	Kilogram
KNN	K-Nearest Neighbors
kPa	Kilopascal
L	Liter
LDA	Linear Discriminant Analysis
LightGBM	Light Gradient Boosting Machine
LR	Logistic Regression
LSTM	Long Short-Term Memory
MAP	Mean Arterial Pressure
meanbp	Mean Blood Pressure
MEWS	Modified Early Warning Score
MIMIC	Mart for Intensive Care
min	Minute
ml	Milliliter
ml/kg/hour	Milliliters per kilogram per hour
MLP	Multi-Layer Perceptron
mm	Millimeter
mM	Millimolar
mmHg	Millimeters of mercury
mmol	Millimole
MSE	Mean Square Error

NA	Not Available
NB	Naive Bayes
NEWS	National Early Warning Score
NICU	Neonatal Intensive Care Unit
NN	Neural Network
NNET	Neural Network
NPV	Negative Predictive Value
NPV	Negative Predictive Value
PCA	Principal Component Analysis
pH	potential of Hydrogen
PLR	Passive Leg Raising
PLS	Partial Least Squares
PMI	Pointwise Mutual Information
PPCA	Probabilistic Principal Component Analysis
PPV	Positive Predictive Value
qsofa	quick Sequential Organ Failure Assessment
RCT	Randomized Controlled Trials
ReLU	Rectified Linear Unit
resprate	Respiratory Rate
RF	Random Forest
RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
RoS	Risk of Sepsis
SAKI	Sepsis-induced Acute Kidney Injury
SAPS	Simplified Acute Physiology Score
SCCM	Society of Critical Care Medicine

SD	Standard Deviation
SHAP	SHapley Additive exPlanations
SIRS	Systemic Inflammatory Response Syndrome
SMOTE	Synthetic Minority Over-sampling Technique
SOFA	Sequential Organ Failure Assessment
SVM	Support Vector Machine
sysbp	Systolic Blood Pressure
tanh	Hyperbolic tangent
TCN	Temporal Convolutional Network
temp	Body Temperature
TN	True Negatives
TP	True Positives
TTE	Transthoracic Echocardiography
US	United States
WBC	White blood cell count
XGBoost	Extreme Gradient Boosting
XOR	Exclusive OR
µg	Microgram
µl	Microliter

1. INTRODUCTION

The pathological condition recognized in contemporary medical discourse as sepsis is characterized by a formidable and dynamic interplay involving the infiltration of microbial pathogens into anatomical compartments typically deemed sterile. This intricate pathophysiological process unfolds within the host organism, resulting in systemic perturbations that are often profound. The deleterious consequences of sepsis are unequivocally underscored by its association with a substantial and worrisome mortality rate, primarily attributed to its predilection for targeting vital organs, most notably the lungs and kidneys, thus compromising their integral functions [1, 2]. This complex clinical scenario necessitates a vigilant and comprehensive approach to both its understanding and management.

Emphasizing the critical nature of expeditious intervention, it is imperative to recognize that the urgency with which sepsis is addressed forms the crux of effective therapeutic outcomes. Prolonged exposure to the pathological milieu perpetuates a precarious cycle, heightening the risk of organ dysfunction, and fostering subsequent harm to the host [3]. The inexorable progression of sepsis, if left unchecked, culminates in a fatality that can be averted through the timely recognition and judicious management of this condition. It is thus imperative to underscore the pivotal role of early diagnosis and treatment, not only as a potential lifesaver but also as a profound modulator of clinical outcomes, significantly augmenting the likelihood of survival and concomitantly mitigating mortality [4].

Sepsis, a medical affliction of global significance, affects an estimated 30 million individuals worldwide [5]. Beyond its acute morbidity and mortality implications, sepsis predisposes patients to an elevated risk of chronic illnesses, enduring neurological impairments, and persistent organ insufficiency. In light of these daunting prospects, the urgency of immediate therapeutic engagement becomes all the more pronounced, further underscoring the potential value of early warning systems in attenuating the severity of this formidable adversary [6].

The clinical presentation of sepsis, a diagnostic conundrum in itself, is marked by a multifaceted constellation of symptoms. These symptoms, while manifesting with a certain consistency across affected individuals, are often marred by their nonspecific nature, compounding the challenge of prompt and accurate diagnosis. Clinical manifestations may encompass fever, chills, tachypnea, cognitive disturbances, and hypotension, a spectrum that, though suggestive of sepsis, is fraught with overlaps with other medical conditions, such as influenza, and demonstrates considerable inter-individual variation [7]. This diagnostic dilemma underscores the critical need for the development of more precise diagnostic tools and strategies, especially those capable of distinguishing sepsis from its clinical mimics.

Crucially, the temporal dimension of sepsis management is well-documented that every hour of treatment delay is inexorably associated with a substantial escalation in mortality, with estimates indicating an alarming 4-8% increase in mortality for each hour of procrastination [8]. Recognizing the imperative of timely intervention, computational methodologies, particularly machine learning algorithms, have emerged as promising tools for expediting the diagnosis of sepsis [9]. These algorithms have the capacity to provide real-time prognostications, potentially affording healthcare practitioners a critical temporal advantage of up to 24 hours in predicting the clinical manifestation of sepsis. This innovative approach not only stands to facilitate the prompt initiation of therapeutic regimens but also underscores the pivotal role that technology can play in the battle against this disease. As such, the integration of machine learning algorithms into the clinical realm holds promise as a key component of future strategies to enhance sepsis management.

In the realm of clinical medicine, a plethora of scoring systems has been devised with the primary objective of gauging the severity of sepsis, a critical undertaking in the effective management of this multifaceted medical condition. These scoring systems, distinguished by their unique methodologies and attributes, hold significant clinical relevance. Among the noteworthy scoring systems employed in clinical practice are the Sequential Organ Failure Assessment (SOFA), quick Sequential Organ Failure Assessment (qSOFA), Modified Early Warning Score (MEWS), National Early Warning Score (NEWS), Acute

Physiology and Chronic Health Evaluation (APACHE) II, Systemic Inflammatory Response Syndrome (SIRS), and Simplified Acute Physiology Score (SAPS) II [10]. These scoring systems leverage various patient characteristics and clinical parameters to facilitate the calculation of a sepsis disease score, thereby aiding healthcare practitioners in stratifying patients based on the severity of their condition.

Each of these scoring systems contributes distinct nuances to the assessment of sepsis severity, offering clinicians a multifaceted perspective on the patient's clinical condition. The quick Sequential Organ Failure Assessment (qSOFA), for instance, is valued for its simplicity and ease of use, with a focus on readily observable clinical signs such as altered mental status, elevated respiratory rate, and hypotension. In contrast, the Modified Early Warning Score (MEWS) and National Early Warning Score (NEWS) emphasize the monitoring of vital signs, enabling early recognition of clinical deterioration [11]. The Acute Physiology and Chronic Health Evaluation (APACHE) II scoring system, on the other hand, takes a comprehensive approach, considering a wide array of physiological parameters to assess disease severity. Systemic Inflammatory Response Syndrome (SIRS) criteria are designed to detect systemic inflammation, while the Simplified Acute Physiology Score (SAPS) II integrates both physiological and chronic health variables for a comprehensive evaluation of disease severity.

The Sequential Organ Failure Assessment (SOFA), in particular, has garnered widespread recognition and acceptance within the medical community as a valuable instrument for identifying patients at risk of sepsis-related mortality. Its endorsement by authoritative bodies such as the Society of Critical Care Medicine (SCCM) and the European Society of Intensive Care Medicine (ESICM) underscores its significance in clinical practice [12, 13]. The utilization of SOFA as a validated tool in the realm of sepsis management speaks to the imperative of accurate and expeditious risk assessment in order to optimize patient outcomes [10, 14, and 15].

By affording healthcare providers valuable tools for the expeditious and accurate assessment of sepsis severity, these scoring systems play a pivotal role in facilitating

timely therapeutic interventions, thereby holding the potential to yield more favorable patient outcomes. The sustained refinement and seamless integration of these scoring systems into the clinical domain underscore the steadfast commitment of the medical community to advancing sepsis management strategies, in turn, underpinning the overarching goal of enhancing patient care [16].

Sepsis, in recent times, has ascended to prominence as a research focal point, primarily owing to its status as one of the leading causes of mortality within the confines of the intensive care unit (ICU). It is sobering to note that sepsis, in its virulent impact, surpasses the annual mortality figures attributed to a confluence of formidable, namely AIDS (Acquired Immune Deficiency Syndrome), prostate cancer, and breast cancer, when considered collectively [17]. This alarming statistic underscores the gravity of sepsis as a medical condition and accentuates its profound societal and clinical implications.

Researchers have responded to this healthcare challenge by embracing the synergy of contemporary technological advancements, notably in the realms of machine learning, deep learning, and artificial intelligence. These cutting-edge tools have been harnessed synergistically with sepsis investigations, affording researchers an enhanced understanding of the multifaceted nature of this condition. The fusion of these state-of-the-art technologies with sepsis research has yielded promising outcomes, most notably the development of predictive models capable of identifying patients at heightened susceptibility to sepsis. This proactive approach not only facilitates earlier intervention but also holds the promise of significantly improving patient outcomes.

The field of machine learning, has emerged as a powerful ally in the quest to unravel the complexities of sepsis. By leveraging vast datasets and intricate algorithms, machine learning models have demonstrated the potential to discern subtle patterns and associations within patient data that may elude conventional diagnostic methods. Such models can analyze an array of clinical parameters, identify key risk factors, and generate prognostic insights that empower healthcare providers with a proactive stance against

sepsis. This paradigm shift in early detection and risk stratification reflects a seminal milestone in the ongoing battle against this disease.

In conclusion, the integration of scoring systems into clinical practice represents a pivotal stride towards enhancing sepsis management. Concurrently, the research focus on sepsis, coupled with the incorporation of cutting-edge technologies, embodies a collective commitment to combat this condition. The symbiosis of advanced analytics and medical researches foster a future wherein sepsis may be met with increasingly effective preemptive measures, ultimately safeguarding the lives of countless individuals.

2. LITERATURE REVIEW

The study under review [18] aims to evaluate the potential of harnessing machine learning approaches on transthoracic echocardiography (TTE) data to anticipate fluid responsiveness in critically ill patients. This undertaking is especially pertinent given the impracticability of using passive leg raising (PLR) in certain scenarios due to limitations in patient mobility. The examination comprises a sample of 100 critically ill patients grappling with severe septic shock and sepsis. Spanning a period of 24 months, this research was orchestrated in the ICU of Narbonne Hospital, France. Several machine learning strategies, such as CART, PLS, NNET, and LDA, were deployed to discern the possibility of predicting fluid reactivity from changes observed in cardiac ventricles through cardiac ultrasound. To gauge the predictive acumen of these methodologies, the AUC was employed. Notably, the PLS model pinpointed key echocardiographic parameters integral for fluid responsiveness prediction. Comparative findings illuminated that the machine learning models for fluid response prediction mirrored the hemodynamic shifts observed during PLR. For patient inclusion, the study mandated that all volume challenges were executed based on the supervising physician's judgment, given the existence of at least one clinical signal of insufficient global perfusion due to sepsis. These clinical markers were defined as presumed infection coupled with signs of systemic dysfunction, verified by a SOFA score of 2 or more. Symptoms included systolic pressure lower than 90 mm Hg, reliance on vasoactive drugs, tachycardia (heart rate surpassing 100 bpm), oliguria (urine output under 0.5 ml/kg/hour for more than 2 hours), mottled skin, prolonged capillary refill, and continuous lactic acidosis above 2 mM. Regarding AUC values, the PLR, CART, PLS, NNET, and LDA models showcased values of 0.76, 0.83, 0.97, 0.93, and 0.90, respectively, while the sample yielded AUC results of 0.77 for PLR, 0.68 for CART, 0.83 for both PLS and NNET, and 0.85 for LDA. In essence, the application of machine learning in this context birthed several models with predictive prowess for fluid responsiveness analogous to the hemodynamic alterations noted during PLR.

In [19], researchers seek to devise a technological solution for the early prediction of sepsis in the broader patient demographic. The investigation conducted an analytical retrospective examination of individuals aged 18 and above admitted to the ICU,

specifically those who displayed no sepsis symptoms upon admission. Central to this research was the development of an algorithm termed "InSight". Impressively, this tool demonstrated a sensitivity of 0.90 and specificity of 0.81 when tested on a novel patient cohort. Its predictive prowess was assessed by determining the AUC, which recorded a value of 0.83 for intervals leading up to three hours before a sustained SIRS event manifested. A critical insight from this study was the collective significance of multiple risk factors in prognosticating a patient's long-term sepsis risk compared to individual risk factors in isolation. The data suggests that by monitoring nine routinely assessed vital signs, it is feasible to foresee sepsis onset with a preparatory window of at least three hours before SIRS symptoms appear within the initial five-hour timeframe. This methodology displays marked superiority over conventional sepsis prediction practices. Drawing data from the MIMIC II Database (Version 3) from 2001 to 2008, the study involved a retrospective assessment of adults admitted to the medical ICU. These patients, crucially, did not meet the SIRS criteria either at admission or during the initial four hours of hospitalization. Moreover, these individuals had comprehensive records of nine specific vital parameters. For effective time-series analysis, their ICU stay data was segmented into discrete hourly intervals, and any missing data was substituted with the nearest previous observation. The research adopted a dual-criteria binary classification method to discern in-hospital sepsis cases. Initially, the patient's records had to exhibit an ICD-9 code indicating a sepsis diagnosis during their hospital stay. Additionally, the patient had to match the 1991-established SIRS criteria, indicative of sepsis and lasting a minimum of five hours. The commencement of the inaugural SIRS event that persisted for five hours was marked as the zero hour. Only adult patients admitted to the medical ICU and not meeting SIRS criteria at admission, with complete records of the nine vital parameters, were considered. Out of these, 1394 met the set criteria, with 159 individuals aligning with the gold standard for sepsis diagnosis. These subjects were then divided randomly into four separate, non-overlapping subsets, paving the way for a 4-fold cross-validation approach in training and evaluating the prediction model. A causal time-series approach was adopted to analyze patient data, focusing on discerning patterns in various health indicators. Nine particular metrics were examined, chosen due to their clinical relevance to sepsis and regular assessment in clinical settings. The data was scrutinized using a sliding observation window spanning five hours. Measurements were classified based on median thresholds and labeled as rising, falling, or static. The study delved into the correlations amongst pairs and sets of three measurements to understand the

interconnectedness of bodily systems, crucial for early sepsis detection. A unique dimensionless score similar to the MEWS was devised by integrating measurement trends, aiming to anticipate sepsis onset. The score's calculations, denoted by p(~), correlated with the probability of specific measurements triggering sepsis onset. This score was calibrated using a standard optimization technique, with calibration constants typically ranging between [0,2]. In essence, the study introduced the "InSight" computational method, designed to predict sepsis onset in ICU-admitted patients. Utilizing nine routinely recorded clinical variables, InSight can predict sepsis onset three hours before a sustained SIRS incident in patients. Its sensitivity of 90% and specificity of 81% overshadows current biomarker detection methodologies. InSight's capability for early, accurate sepsis identification can potentially expedite interventions, streamline antibiotic administration, and possibly reduce associated complications and extended hospitalizations. The algorithm's strength lies in amalgamating diverse measurements and discerning connections between these measurements and vital patient outcomes. Despite its promise, the study's retrospective nature and reliance on a limited clinical metric set presented constraints. Nevertheless, InSight emerges as a potentially transformative clinical tool for sepsis trajectory prediction.

In the research [20] explorations, the detection of sepsis, a life-threatening condition resulting from infections, remains a pivotal concern. Sepsis detection is fundamentally linked to discerning alterations in physiological indicators and symptoms of the body, echoing earlier academic studies. Specifically, the emphasis has been on the window known as the "target detection time" – typically characterized by the manifestation of at least two SIRS criteria within the first hour of a confirmed sepsis infection. The present investigation, as delineated in the research, utilized ICD-9 code 995.9 from the ninth revision of the International Classification of Diseases to denote a hospitalization episode. Furthermore, in alignment with past academic works, the researchers employed the "early" detection criterion, essentially spanning three hours before the identified detection hour. Utilizing the expansive MIMIC II database, the study engaged in a retrospective examination. The criteria for inclusion in the study were patients aged 18 or older who had recently been admitted to a medical ICU and exhibited SIRS symptoms either within their first ICU hour or within the initial four hours of hospitalization. The research pursued a two-fold strategy: firstly, constructing deep learning models with the ability to

identify initial sepsis indications. Secondly, comparing these deep learning models with the InSight regression model, which employs conventional temporal feature extraction methods. The comparison revealed that deep feedforward networks surpassed the InSight model's performance, as attested by the AUC scores of 0.887 and 0.915. Remarkably, when amalgamating both feature sets, the AUC value remained consistent at 0.915, mirroring that of the primary feature set. The LSTM model, using the fundamental feature set, registered its peak AUC at 0.929. An intriguing aspect of this study is its deliberate eschewal of domain-specific feature extraction. By directly comparing with reference features, the research underscored the efficacy of artificial neural networks in autonomously gleaning significant features. Moreover, the findings highlighted the superior capability of feedforward neural networks integrated with LSTM in extracting meaningful patterns. Delving into the specifics, the researchers curated nine essential parameters from the MIMIC II database, ranging from blood oxygen saturation, age, heart rate, to white blood cell count. Two distinct datasets were derived from these parameters: one encompassing essential summary data and the other constituting features previously utilized in the InSight model. To address inconsistencies in the timing of individual parameter measurements within electronic medical records, the team synthesized patient data by computing hourly metrics such as minimum, mean, and maximum values. This hourly interval was a conscious choice, reflecting the prevalent frequency for each variable. In the absence of specific measurements, proximate recorded values were used as substitutes. The research also laid out a meticulous method for gleaning reference features from a five-hour window, focusing on discerning any data value shift. This method encapsulated calculations of averages within stipulated intervals, subsequent differentiation, and categorization of these variations. Crucially, correlations between dual and triad measurements were computed, excluding age as a determinant. In this study, two DNN architectures were explored: deep feedforward neural networks (or multilayer perceptrons) and LSTM networks, with the latter being celebrated for capturing long-term dependencies. Notably, LSTM networks leverage memory blocks with in-built gates to regulate and store selective information. The primary intent of the research was to appraise the capability of deep learning models in sepsis detection vis-à-vis the InSight model. The scholars examined three deep learning models anchored in diverse feature sets. The ultimate results indicated the superiority of the SepLSTM model, which exclusively employed fundamental features, over the InSight model, substantiated by its AUC of 0.929. When juxtaposing the specificity and sensitivity metrics of the

developed models against InSight and SIRS, SepLSTM exhibited a heightened specificity, albeit with a slight dip in sensitivity. In conclusion, this seminal study offers a robust argument in favor of deep learning models, emphasizing their inherent prowess in sepsis detection without resorting to manual feature extraction, thus marking a significant stride in the domain of medical research.

In the quest to enhance the predictability of sepsis onset, a comprehensive study detailed in [21] the efficiencies of RNNs against the InSight algorithm. The research revolved around adult patients in ICU settings without sepsis, as deduced from the MIMIC III database. Using the AUC as a measuring metric, findings showcased the InSight algorithm securing an AUC of 0.72, while the RNN exhibited an AUC of 0.81 for the three-hour onset of sepsis prognosis. Moreover, at a sensitivity threshold of 90%, InSight registered a specificity of 31.1%, whereas the RNN reached to 47.0%. Beyond the immediate prognosis, the study also ventured into extended timeframes, assessing predictability across 6 and 12-hour intervals. Drawing conclusions, the RNN's superior performance overshadowed InSight's predictive capacities. Researchers posited the potential of RNNs in forecasting ICU-based sepsis occurrences. To differentiate between participants, a division was established depending on whether sepsis was contracted during hospitalization, utilizing diagnostic criteria involving ICD codes and a 5-hour SIRS interval. A distinctive feature of this research was its dive into the uncharted waters of the ramifications of interpolations on sepsis onset discernment. Deriving data from the MIMIC III repository, the investigation encapsulated patient data. From this, 18 distinct datasets were curated, factoring in 6 interpolation levels for sepsis onset detection and three prediction timeframes. Guided by the methodologies outlined by Calvert et al., parameters were extracted, with hourly means for each computed. Addressing data lacunae, the researchers employed dual methods: linear interpolation coupled with "carry forward/backward" for gold standard implementation and a "carry forward" method for classification tasks. This investigation shed light on the InSight algorithm's prowess in extracting an expansive 101-feature set from the lookback. Feature extraction spanned mean computations, parameter value shifts, and complex parameter combinations, providing insights into the propensity for sepsis based on singular or combined parameter values. A pivotal aspect of the study was the leveraging of RNNs to discern sepsis onset by capitalizing on evolving temporal patterns. The RNN architecture encompassed two

hidden layers, each fortified with 40 neurons and underpinned by Gated Recurrent Units (GRUs). Through binary cross-entropy and the Adam algorithm, network optimization was achieved. The RNN's objective was dual-fold: aiding healthcare professionals with timely alerts and allowing clinicians the final say. The research's crux was the comparative analysis between RNN and InSight, where RNN consistently outperformed InSight, particularly evident in extended vital sign sequences. However, it's worth noting the AUC, sensitivity, and specificity values, although below benchmarks set by analogous studies, showcased a more generalized classifier due to a larger dataset. Such revelations beckon further inquiry into sepsis onset discernment, focusing on data lacunae management and variations in SIRS interval durations. Machine learning efficacy in sepsis prediction extends beyond the classifier, interlinking with data quality and the gold standard. The intricate task of pinpointing sepsis onset is riddled with challenges, especially when prolonged borderline symptoms confound the diagnosis. The research acknowledges inherent biases in the MIMIC III database, suggesting neural network fine-tuning with nuanced data sets as a mitigation strategy. RNNs, despite their opaque nature, hold potential as a supplementary clinical tool, but the need to enhance specificity to reduce false alarms is paramount. The study further highlights the challenges of classifier deployment beyond ICU settings. The overarching vision is the amplification of sepsis prediction precision through iterative advancements in machine learning methodologies applied to dynamic electronic health record data.

In the study [22], a two-stage framework named HeMA is developed, leveraging machine learning algorithms for the early prediction of sepsis onset. At its core, HeMA synergistically integrates machine learning models in its initial stage, followed by statistical tests in its subsequent stage. When evaluated on datasets comprising different proportions of sepsis cases (50% and 25%), HeMA showcases significant advancements in specificity and precision. The empirical evaluation of this research primarily utilized patient physiological data sourced from the Cerner CareAware iBus® platform, within a comprehensive hospital setup located in the Southeastern United States, which boasts six ICUs. Core physiological metrics, such as heart rate, blood pressure, respiratory rate, and oxygen saturation levels were meticulously recorded. The widely accepted Sepsis-3 definition was employed to identify potential sepsis cases and ascertain the optimal time window for sepsis onset. Delineating their data collection process, researchers gathered

data from patients over two distinct observational periods- six hours and ten hours prior to sepsis commencement, leading to two individual datasets. The presented HeMA framework follows an advanced hierarchical approach for sepsis detection. In its design, two machine learning models (RF and NN) leverage physiological data to ascertain the likelihood of sepsis in a patient, and subsequent statistical tests help finalize the decision. A noteworthy aspect of the HeMA model is its three sub-stages in the second stage, where decisions pivot around the probabilistic outcomes from the first stage. This involves creating baseline probability distributions, applying the Kolmogorov-Smirnov test for pattern anomalies, and subsequent decision-making based on refined p-values. The intricacies of the first-stage RF model, as well as the NN structure, are elaborated upon, revealing the robustness of their approach. In benchmarking the HeMA framework, the authors leverage the first-stage model outputs as critical reference points. By applying a 0.5 threshold to the first-stage probabilities, distinctions are made between sepsis and non-sepsis cases. This threshold manipulation facilitates an exploration into the balance between model sensitivity and specificity. While there's an evident upswing in model performance, certain trade-offs, such as slightly reduced sensitivity, are apparent. However, like any academic research, this study isn't without limitations. While the framework augments specificity and positive predictive value, it marginally compromises sensitivity, implying potential challenges in identifying sepsis patients. The retrospective nature of the study, the need for model retraining across varied datasets, and the limited physiological data streams examined necessitate further exploratory endeavors. In conclusion, while promising, the HeMA framework requires meticulous refinements and potentially pilot studies before its full-scale clinical deployment.

The academic research [23] is delved into the application of machine learning for early sepsis prediction, a critical area with significant clinical implications. One approach harnesses physiological data available in digital health records to create a predictive framework. Central to this method is the pointwise mutual information (PMI) matrix, adept at identifying both linear and non-linear correlations among clinical covariates within Electronic Health Records (EHRs). For records of a particular stay duration, length (L), this method combines PMI matrices horizontally to form a 3-way tensor. Subsequently, using Tucker decomposition, these tensors are deconstructed to retain core tensors, which then serve as the foundational feature set for predicting sepsis onset.

Notably, this process leverages light gradient boosting for binary classification and has demonstrated promising outcomes. The LightGBM model in particular addresses the challenge of class imbalances commonly found in sepsis datasets. The method's efficacy is exemplified through its performance on the PhysioNet/Computing in Cardiology Challenge 2019 dataset. Metrics such as the average normalized utility score and the AUC underscore its superior predictive capabilities. The dataset consists of electronic health records from ICU patients across three healthcare institutions (A, B, and C), totaling 64155 patients. The methodology utilizes forward fill preprocessing for EHR data imputation. The statistical measure, PMI, plays a pivotal role in understanding the correlation between two random variables, and its applicability extends to assessing the association between clinical covariates and sepsis onset within a six-hour window. PMI aids in identifying the covariates with the highest predictive potential. This PMI-based approach delineates the difference between the common probability distribution and the product of their individual probabilities. The PMI matrix, thereby, provides insights into temporal interactions between clinical values. In essence, high PMI values suggest frequent interactions, while lower values indicate the opposite. The study's tensor-factorization method, which employs the PMI matrices, captures the multifaceted relationships among patient predictors post-ICU admission. After decomposing the tensor, it surfaces patterns and relationships pertinent to sepsis onset. The extraction of 120 latent features from patient-tensors, which elucidate tripartite temporal interactions between covariates, is an integral part of this process. The study indicates that machine learning models, especially when compared to traditional measures like SIRS, qSOFA, and MEWS, can offer superior accuracy in predicting sepsis onset with minimal clinical parameters. To assess the results, a custom model that either rewards or penalizes predictions within specific timeframes from sepsis onset was deployed. The comparative analysis against baseline models revealed the superiority of tensor factorization methods, with the core tensor accounting for intricate interactions between tensor components leading to enhanced classification outcomes. The study effectively employs the PMI matrix to gather pairwise clinical associations from chosen covariates, leading to improved sepsis detection outcomes.

In the meta-analysis [24], the authors seek to contrast the proficiency of machine learning models in forecasting the onset of sepsis against that of traditional scoring methodologies.

An exhaustive review of the literature led to the selection of seven pertinent studies that satisfied the predetermined eligibility criteria. These machine learning algorithms demonstrated an aggregate AUC of 0.890 when tasked with predicting sepsis onset within a window of 3-4 hours. These models attained a sensitivity and specificity of 0.810 and 0.720, respectively. On the other hand, the combined AUC for traditional evaluation tools, including SIRS, MEWS, and SOFA, spanned between 0.500 and 0.780. Notably, the diagnostic odds ratio for machine learning stood at 15.170, which surpassed those of SIRS, MEWS, and SOFA. Hence, this research underlines the superior efficacy of machine learning models in forecasting sepsis onset compared to established scoring systems.

In the academic research documented in study [25], the primary objective is the development and assessment of innovative sepsis diagnostic tools employing machine learning algorithms. These tools were juxtaposed with traditional diagnostic practices. The research centered on adult patients who accessed care in an emergency room. Sourcing retrospective electronic health record data from a singular medical center, the study incorporated triage details comprising health metrics, preliminary features, and primary general concerns. Four distinct machine learning methods—neural network, logistic regression, gradient boosting, and random forest—formed the basis of the research. Results underscored the superior diagnostic capability of all these machine learning models over standard benchmarks such as qSOFA, MEWS, and SIRS. The research timeline spanned 24 months, commencing in June 2018 and concluding in May 2020. Conducted within an urban academic hospital's emergency department, all adult individuals seeking emergency medical intervention were considered, barring those who departed prematurely, those who chose to leave against medical advice, and individuals deceased upon arrival. Diving deeper into the research methodology, a blend of structured and unstructured datasets was employed to prognosticate sepsis onset in emergency care. While the structured dataset encompassed vitals, demographics, arrival means, urgency level, and health histories, the unstructured component was based on triage nurses' notes recorded in Thai, elucidating the primary concerns of the patients. Eight models, spanning four machine learning techniques, underwent performance comparison, particularly emphasizing the influence of textual data incorporation. Traditional scoring systems like qSOFA, SIRS, and MEWS were set against machine learning tools with diagnostic

thresholds set at qSOFA: 2, MEWS: 5, and SIRS: 2. For feature engineering and data preprocessing, techniques like one-hot encoding, text extraction, tokenization, and TF-IDF transformations were applied. To address the imbalanced nature of sepsis diagnoses, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented. Performance metrics included AUC, classification matrices, and AUPRC. Of the 133707 emergency room visits analyzed, eight machine learning models were developed for sepsis forecasting. The standout model amalgamated triage specifics, patient demographics, and primary symptoms, attaining an impressive 0.931 AUC and 86.940% sensitivity. Particularly, these algorithms excel in sepsis prediction utilizing initial clinical information and free text from triage spaces. Even with inherent limitations, the study epitomizes the potential of localized machine learning models in early sepsis detection, facilitating timely clinical interventions.

The study [26] delves into a novel approach to monitor the clinical signs of sepsis and the state of organ systems in an ICU setting without the conventional lab tests. At the heart of this methodology lies the Deep SOFA-Sepsis Prediction Algorithm (DSPA). DSPA ingeniously amalgamates features extracted from Convolutional Neural Networks (CNN) and integrates them with the Random Forest (RF) algorithm. The ultimate objective is to provide a robust predictive mechanism for SOFA scores in patients' potentially developing sepsis. When tested against a specific section of the MIMIC III dataset, the results were promising. Notably, the DSPA produced significant metrics: a correlation coefficient of 0.863, a mean absolute error of 0.659, and a root mean square error of 1.230. Impressively, the DSPA surpassed traditional machine learning and deep learning techniques, producing an AUC of 0.982 for imminent sepsis and 0.972 for a six-hour pre-sepsis prediction. This venture extracted vital signs from ICU patients, courtesy of the MIMIC III dataset, providing a foundation to track sepsis indicators and the overall health of organ systems. The methodology was underpinned by rigorous processes, including data preprocessing, training, testing, and handling missing data. The dataset was derived from the expansive MIMIC III database, incorporating data from over 53000 patients. For this study, a more focused subset, "Sepsis-3 in MIMIC III," was employed, resulting in a final sample of 11791 patients. Twelve hours preceding a predicted sepsis onset, data on seven distinct vital signs were recorded. Challenges like variable data frequency intervals were addressed using data imputation strategies. Among the gamut of techniques

available, the PPCA method was chosen due to its efficacy and superior performance in representing missing values. Delving deeper, the model adopted CNN for feature extraction and RF for tapping into data interrelationships. The study leveraged ten-fold cross-validation and a suite of metrics to ensure the robustness of the models. In essence, the DSPA algorithm stands out as an avant-garde deep learning estimation tool to prognosticate SOFA scores, ultimately predicting sepsis onset and gauging organ failure severity in ICU settings.

In a comprehensive research endeavor [27], a machine learning based sepsis screening tool is developed and evaluated. The study analyzed electronic health record data retrospectively from 2759529 adult patients who visited 49 urban community hospitals' emergency departments over a span of 22 months. When assessed, this screening tool, termed the Risk of Sepsis (RoS) score, showcased an impressive AUC ranging from 0.93 to 0.970. The research underscores the necessity for external validation of the RoS score at other independent sites. The patient data in this research came from adults aged 18 and above who visited the emergency departments of Tenet Healthcare. As part of the assessment, in-hospital mortality served as a secondary outcome to gauge the RoS score's effectiveness. Features for the model were gleaned from reviewing extant models, expert consultations, and supervised machine learning. This comprehensive approach allowed for a feature set inclusive of 56 potential model inputs. The study revealed a compelling approach to feature engineering that's effective for predicting sepsis. The team managed missing data by imputing them with extreme values (-9,999) and trained a gradient-boosted model comprising myriad decision trees. Post iterative feature selection, the RoS score model encapsulated 13 pivotal attributes, with lactic acid being paramount. The model's performance metrics, AUC, sensitivity, specificity, precision, and alert rate, were calculated. Performance metrics were subsequently evaluated at periodic intervals following the index time to ensure model pertinence. Its superior discrimination was evident across all time intervals, demonstrating significant advantages over benchmarks like SOFA and SIRS. Despite its advantages, the RoS score did exhibit an inclination to underestimate sepsis occurrences, suggesting areas for refinement. While the RoS score's creation relied on a machine-learning model trained on clinical sepsis criteria, it's worth noting the inherent limitations of the study due to the absence of a universally accepted clinical standard for sepsis. Although machine learning models can often seem opaque,

the authors posited their appropriateness for clinical implementation. Established benchmarks like SIRS have been in use for over two decades, making the shift to newer tools like the RoS score potentially challenging. The RoS score's high false-positive rate, leading to potential over-treatment, further reinforces the need for future research to fine-tune its clinical application. Sensitivity analyses conducted without lactic acid results still showcased the RoS score's superior discriminatory capabilities over benchmarks. In conclusion, while the RoS scoring model exemplifies cutting-edge research in sepsis detection, further independent evaluations are essential to validate its effectiveness across various settings.

The research [28] constructs a composite model using clinical information to anticipate the likelihood of sepsis-induced acute kidney injury (SAKI), leveraging a stacking algorithm composed of four base-level machine learning methods. The results showcased notable generalizability across different databases, spanning both the US and China. The model's compatibility with multi-center external datasets and its user-friendliness for clinical practitioners were highlighted. Sepsis, an adverse response to infection, is lethal, accounting for over 30% of fatalities among ICU patients. A prevalent consequence of sepsis is organ failure, with the kidneys often being most affected. Over half of the ICU cases of acute kidney injury (AKI) due to sepsis are linked with chronic kidney disease and elevated long-term mortality. Prompt diagnosis and intervention of SAKI are vital to enhance patient prognosis. Given the urgency in identifying and addressing SAKI — as standard diagnostic criteria often detect damage only after it has set in — machine learning has been championed for its potential in early and accurate prediction across diverse datasets. Among various methodologies, ensemble learning has manifested superiority over individual algorithms and conventional regression models. This investigation aimed to cultivate an ensemble model, amalgamating SVM, RF, NN, and XGBoost algorithms, to predict AKI risk in sepsis patients adeptly. The research encompassed data from 45390 patients, 24352 from eICU-CRD, and 21038 from MIMIC IV. This model, built on clinical data, exhibited commendable discriminative capabilities, achieving optimal performance 12 hours prior to AKI emergence. The XGBoost method, having a significant weight in the ensemble, facilitated the creation of an online AKI risk calculator, viable within a 12-hour frame for sepsis patients, which clinicians can access via downloadable materials. The study utilized MIMIC IV data, prioritizing patient

demographics, lab results, and vital signs for SAKI prediction. By simplifying model intricacies via a two-step feature screening approach with the four algorithms, the ensemble model demonstrated noteworthy prediction capabilities. An online risk calculator was crafted using XGBoost to provide clinicians a tool predicting imminent AKI risk in sepsis patients within 12 hours. This model encapsulates complex aspects of AKI's pathogenesis, diagnosis, prognosis, and management. It was found that features such as temperature, heart rate, hemoglobin, and oxygen saturation can be used as potent SAKI predictors. The study also delved into the pathophysiology of SAKI in septic patients, spotlighting the pivotal roles of inflammatory agents and imbalanced infection responses. It underscored the urgency of timely antibiotic administration and infection source eradication to mitigate AKI risks. The research hinted at the multifaceted nature of predictive indicators, suggesting that clinicians should be vigilant of changes in these metrics for proactive SAKI management. The research, being retrospective, poses certain limitations. The ensemble model's reliance on only four machine learning algorithms highlights potential areas for enhancement. Combining datasets can introduce outliers and dilute information depth, underscoring the need for future investigations to consider longitudinal modeling or temporal data summarization.

The significance of early identification and prediction of sepsis within ICU patients cannot be overstated, given its potential to enhance patient prognosis and reduce healthcare expenditure [29]. Drawing from the PhysioNet Challenge 2019 dataset, comprising 40336 patient files across two hospitals, the study underscores the challenges posed by imbalanced data. Specifically, only 2932 out of 40336 patients were diagnosed with sepsis. To amplify the predictive capability, the sepsis label's temporal alignment was pre-empted by six hours across datasets, noting a significant volume of variables showing over 70% missing values. Addressing data imbalance is critical. In this context, the Imbalanced-learn library is recommended, offering resampling techniques like SMOTE analysis. The study promotes techniques like feature extraction and imputation for handling redundant and missing data. Among several methods for addressing missing data, Missforest emerges as superior due to its adaptability across various data types and its ability to maintain a normal data distribution. The study aimed to identify six primary physiological markers for predicting sepsis early in ICU patients. Through correlation analysis, high-contribution features were distilled as potential predictors. Machine

learning algorithms, especially XGBoost and RF, showcased their prowess in predicting sepsis. Notably, the Multilayer Perceptron Neural Network (MLP) demonstrated its ability to be trained on numerical models without any preconceived data distribution notions. Comparatively, the study revealed XGBoost's superior performance in sepsis prediction after implementing optimal feature selection and SMOTE analysis. The models' evaluation metrics included ROC curves, precision-recall curves, and AUC scores. Using hourly ICU patient data, the study aimed to predict sepsis onset. Objectives encompassed predicting prognosis, septic shock levels, and maximum ICU stay duration. Multiple machine learning algorithms like MLP, XGBoost, RF, LightGBM, and LL showcased substantial predictive capabilities. A unique insight from the study highlighted that females might be more susceptible to severe sepsis outcomes due to differential hormonal responses to infections. In conclusion, this comprehensive study accentuates the potential of cutting-edge machine learning algorithms in tackling the intricacies of sepsis prediction in ICUs, considering the complexity of diseases and data imbalance.

In the study [30], researchers are introduced the temporal convolutional network (TCN) as a novel approach to predicting sepsis in ICU-admitted patients who do not initially exhibit sepsis criteria. The intrinsic capability of convolutional networks in discerning temporal patterns underpins its superiority. This research particularly champions a specific variant of CNN, the dilated causal convolutional network (or TCN), for sepsis prediction. Comparative evaluations between the TCN, LSTM, and GRU models reveal that TCN holds a more extended memory, translating to enhanced performance, especially in binary classification tasks using the MIMIC III dataset. To devise accurate sepsis predictions, the researchers leveraged data from the MIMIC III database, which comprises 58976 admissions from 46520 patients. The researchers innovatively recommend substituting RNNs with TCNs to gain insights from historical data. By deploying causal convolutions, they reconfigured the conventional 1D convolutional layer, ensuring predictions are solely based on past data, eliminating potential future data influences. This shift emphasizes the importance of causal and dilated convolutions in refining the TCN model for time series applications. For benchmarking, models such as LSTM, RF, and AdaBoost were utilized. The assessment metrics highlighted not only the TCN's commendable accuracy but also its superior recall, F1 score, and AUC, compared to other established models. Conclusively, this research accentuates the potential of TCN,

alongside other advanced deep learning methodologies, in discerning intricate physiological correlations, with a prime goal of mitigating sepsis-related fatalities in ICU settings. The findings underscore the pivotal role of temporal patterns in sepsis detection and the necessity for data enrichment, especially in deep learning models with relatively limited datasets. Future research avenues might explore attention models as alternatives to TCNs and strategies like focal loss to address data imbalances, thereby enhancing model efficacy.

The referenced research [31] introduces a cutting-edge method for extracting relevant features, drawing attention to the relationship between patient stability and sepsis probability during ICU care. The current investigation leverages machine learning algorithms and ICU bedside monitor data to foresee sepsis onset. By analyzing data from the preceding eight-hour window, these algorithms forecast the potential sepsis risk within the subsequent four hours. The SVM with a radial basis function stands out, boasting an AUC value of 88.380%. The study introduces an innovative technique to construct a predictive model to anticipate sepsis in ICU-admitted adults. Initially, it examines variability across four regularly monitored vital signs. Moreover, the research highlighted that shifts in metabolic parameters, often tracked via Electronic Medical Records (EMRs), could mark sepsis's onset. The research utilized EMRs from ICU patients at the Israeli Rabin Medical Center from 2007 to 2014. They focused on the SIRS, characterized by multiple symptoms. The main aim was to confirm sepsis presence by assessing SIRS criteria alongside a verified infection. Adult participants in the study were 18 or older and spent a minimum of 12 hours in the ICU. Only those with complete data records were considered. Of the entire ICU admissions between 2007 and 2014, about 35.4% (1605 patients) were identified with sepsis-related infections. Out of this, only 401 had exhaustive data sets, and from this pool, a sample of 300 was chosen, focusing on their sepsis diagnosis at the antibiotic administration time. The authors crafted a technique to gauge vital sign fluctuation, recording numerous measurements for each sign prior to predicting sepsis. These measurements were split into two distinct 12-hour periods. Furthermore, they identified five key attributes from each measurement set related to the vital sign extremes. The variability within each vital sign was assessed based on the intensity and frequency of changes. For a more in-depth analysis, they compared their extracted features with Guillen's. The team discerned differences in the behavior of

the Mean Arterial Pressure (MAP) through their unique features. Their research aimed to minimize dimensionality by focusing on four pivotal features. The entire process of feature selection was bifurcated. In the second phase, they settled on a concise model encompassing specific vital sign changes. This study dives deep into sepsis prediction, utilizing various machine learning models. The primary goal was identifying the most efficacious algorithm in terms of AUC. SVM emerged as the top performer, achieving an AUC of 88.380%. The researchers compared their results with previous works and highlighted the unique contributions of their study. However, it's crucial to recognize its limitations, such as the small data set, reliance on antibiotic administration timing as the sole sepsis indicator, and outdated sepsis definitions. Such constraints might impede the broader application of the study's findings.

In the research [32], early sepsis prediction six hours prior is made possible feature generation, and supervised classification algorithms like XGBoost and LightGBM. The feature generation technique harnessed statistical power, window components, and medical attributes to develop models. Miceforest was employed to manage substantial missing data. The study suggests that LightGBM outperforms in terms of generalization and processing speed on multi-dimensional datasets, noting an AUC of 0.979 for the feature generation method. Key risk factors for early sepsis were identified as PTT, WBC, and erythrocytes. Current sepsis scoring systems for clinicians are plagued by false alarms due to a lack of specificity. Challenges in this domain revolve around the utilization of diverse physiological metrics and the creation of effective machine learning algorithms. The research underscores the utilization of AI in diagnosing, prognosing, and treating sepsis. Two primary machine learning techniques highlighted are supervised learning and reinforcement learning. Input variables for these models comprise physiological markers, biomarkers, laboratory outcomes, and demographic information, but missing values in these datasets present a challenge. Machine learning techniques, including SVM, XGBoost, RF, lasso regression, and NN, are discussed in the context of sepsis prognosis and diagnosis. Despite the demonstrated success of several studies in predicting sepsis, there remain gaps in explanation and generalization capacities of machine learning techniques in this domain. Using a SHAP value metric, the study aimed for model transparency. Data preprocessing methods based on XGBoost and LightGBM were devised to assist early sepsis detection. By examining the relationship between model

predictability and class imbalance, the study aimed to develop models with clinical interpretability and generalizability. The study employed a comprehensive dataset from three hospitals with physiological ICU data on 22336 patients, 1714 of whom had sepsis. Challenges like missing values and class imbalance were addressed using mean processing and feature generation. Multiple imputation techniques, especially Miceforest—a technique derived from the RF chain equation—are preferred over single imputation when faced with substantial missing data. The study utilized two tree algorithms, XGBoost and LightGBM, to predict outcomes, with metrics such as precision, recall, and F1-score employed for evaluation. After handling class imbalance and missing values, the dataset was segmented temporally for statistical metric extraction. New medical diagnostic metrics like shock and oxygenation indices were added to enrich the model. Post data processing, a dataset of 23711 physiological data points was obtained. Out of 25 selected variables, those with over 98% missing values and certain demographic indicators were excluded for focusing on early sepsis patterns in physiological data. The study's conclusion underlined the superior performance of the feature generation method trained with LightGBM. The Miceforest algorithm was acknowledged for effectively handling missing data. The conventional mean processing method's limitations were highlighted, though improved mean processing showed promise. LightGBM was favored over XGBoost for handling extensive data due to its memory efficiency. The research advocated for more research on imbalanced data handling and emphasized future exploration into new variables for enhanced prediction accuracy.

Sepsis remains a significant challenge in healthcare due to its high mortality and costly treatment, as elaborated in reference [33]. A spectrum of models was assessed in this study, spanning RF, LR, SVM, NB, ensemble techniques, and a pioneering ensemble approach. Using clinical test data and health indicators, the goal was to craft a machine learning system adept at predicting and detecting sepsis among ICU patients. The newly proposed ensemble approach notably surpassed other models, registering a balanced accuracy of 0.960. Recent advancements have steered sepsis definitions towards symptom-based ICU risk scores rather than solely focusing on infection. This research delves into machine learning's potential within healthcare, spanning disease identification, diagnosis, antibiotic choice, and advanced health monitoring. Employing

models like XGBoost, RF, LR, SVM, and NB, the study harnesses ICU databases, concentrating on clinical lab results and vital signs. The research introduces a machine learning solution adept at rapidly pinpointing and categorizing sepsis in ICU-admitted patients, leveraging matrix operations to optimize performance and imputation techniques for data gaps. This paper presents a model-based machine learning approach using the publicly available Skaraborg Hospital Dataset, with 1572 sepsis-diagnosed patient records from 2011-2012. After data cleaning and transformation, the set was divided into training (1257 records) and testing (315 records) segments. Methods like kernel density estimation for outlier identification and principal component analysis for dimensionality reduction were incorporated. The machine learning models processed essential variables like age, gender, vital signs, SIRS criteria, and various blood parameters, using Python tools for data visualization and model evaluation. The research also examines ensemble learning, which aggregates predictions from individual models to bolster accuracy. Three major ensemble techniques—bagging, boosting, and voting—are detailed, with emphasis on their respective merits and limitations. The introduced framework aspires to detect and predict sepsis onset, encompassing classifiers like SVM, RF, NB, LR, and XGBoost. Public ICU datasets for sepsis cases were harnessed, and evaluation metrics such as confusion matrix, accuracy, and sensitivity were introduced. The ensemble model's efficacy was assessed in comparison to existing binary classification techniques. Sensitivity was deemed crucial for model efficacy. The study leverages various models, from SVM to XGBoost, complemented by an ensemble technique, which amalgamates these models to enhance classification performance. Challenges surrounding incomplete datasets in healthcare are underscored, necessitating advanced imputation methods. Potential model enhancements might involve more comprehensive patient data collection in subsequent versions. The significant data voids in the utilized dataset could hinder model performance, prompting a shift towards sophisticated imputation strategies, potentially elevating model accuracy through statistical techniques or dedicated algorithms.

The study [34] aims to conduct a comprehensive assessment of the extant literature concerning the utilization of Clinical Decision Support Algorithms (CDSAs) that capitalize on non-invasive metrics to predict neonatal sepsis. An exhaustive search across CENTRAL, EMBASE, and PubMed databases yielded 36 studies involving 18096

infants, selected subsequent to rigorous screening and data abstraction processes. Most CDSAs scrutinized in these studies concentrated on heart rate-based metrics. The researchers established that the integration of heart rate-based metrics in CDSAs, particularly when amalgamated with vital statistics and demographic details, presented a reliable methodology for neonatal sepsis prognostication. However, the generalized application of CDSAs in clinical milieus faces certain roadblocks, including an absence of unequivocal evidence and a dearth of standardization protocols outside of controlled research arenas. The study review also delves into the complexities of diagnosing neonatal sepsis, arising from the neonate's immature immune and autonomic control systems. Traditional diagnostic methods, reliant on invasive blood tests and biomarker analyses, may not be optimal for identifying at-risk newborns. According to the assessment, non-invasive metrics such as vital signs show promise in foreseeing life-threatening conditions. Additionally, the review differentiates between conventional algorithms, which are based on human-crafted predictive factors, and machine learning algorithms that make optimal predictions by incorporating a variety of input covariates. The objective of this review is to gauge the effectiveness of utilizing non-invasive vital sign monitoring to predict neonatal sepsis, with a focus on the available evidence supporting the efficacy of CDSAs. Despite the reliability demonstrated by CDSAs—especially those incorporating heart rate-based metrics with vital statistics and demographic information—their large-scale adoption outside of research settings remains problematic, primarily owing to inconclusive evidence and a lack of standardization. To enhance the predictive precision of CDSAs and thereby facilitate their broader clinical acceptance, further investigations are imperative. The efficacy of new CDSAs, their safety profiles, and various parameter combinations must be examined in future randomized controlled trials (RCT) to substantiate their role in neonatal care. The current body of evidence supporting the clinical utility of CDSAs in non-research settings remains insufficient to warrant their mainstream application for predicting neonatal sepsis. Therefore, further RCTs are essential to confirm the viability of heart rate-based and other non-invasive metrics in predicting sepsis, as well as to evaluate the safety and risks associated with the deployment of CDSAs in neonatal healthcare settings.

The primary aim of the study [35] is to enhance early detection of neonatal sepsis by utilizing machine learning algorithms that analyze non-invasive physiological and

demographic data. A cohort of 325 infants, each with unique event histories, was included for analysis. The researchers employed time-domain features extracted from heart rate, respiratory rate, and oxygen saturation measurements, in addition to considering demographic variables. Utilizing the NB algorithm to forecast sepsis, the model's performance was notably enhanced by the incorporation of additional vital signs. The algorithm demonstrated predictive capacity within 24 hours of initial clinical suspicion, as evidenced by an AUC value of 0.820. For the machine learning model's construction, vital signs and EHR data from standard Neonatal Intensive Care Unit (NICU) monitoring systems were employed. The study's overarching objective was to create a universally applicable early sepsis detection framework. The data used for this investigation were collected from neonates admitted to Stockholm NICUs between the years 2016 and 2020. In addition to heart rate variability, the study scrutinized the predictive efficacy of other commonly monitored acute physiological markers, such as respiratory rate and peripheral oxygen saturation levels. The algorithm's proficiency was particularly assessed in the context of neonates with extremely low birth weights and compared to existing algorithms. This research contributes a pioneering methodology for neonatal sepsis detection, showcasing robust predictive capabilities through the integration of frequently monitored vital signs and demographic factors. The results highlight the potential of merging machine learning algorithms with clinical support systems to enhance personalized care, optimize healthcare resource allocation, and reduce both morbidity and mortality rates within the NICU setting. However, the study is not without limitations. The relatively small sample size could have restricted the statistical power to detect false negatives, thereby increasing the likelihood of discovering false positives. Furthermore, certain variables such as the extent of respiratory support provided and the fractional delivery of oxygen were not considered, factors which could potentially impact the performance estimates of the algorithm. Given these constraints, there is a pressing need for further research to refine the algorithm's performance. To achieve more robust and reliable results, the model should be validated on a larger and more diverse sample of neonates.

The retrospective case-control study [36] aims to formulate machine learning models utilizing EHR data for the early detection of infant sepsis—specifically, four hours before clinical diagnosis. The research focused on neonates admitted to the NICU of the

Children's Hospital of Philadelphia between September 2014 and November 2017. Inclusion criteria stipulated a minimum NICU stay of 48 hours and completion of at least one sepsis evaluation before reaching 12 months of age. The study considered two outcomes for sepsis evaluation: culture-positive sepsis and clinically diagnosed sepsis. The analytical dataset comprised a 44-hour period leading to the sepsis evaluation, deliberately excluding the final four hours. A set of 36 features from EHR data was harnessed for the predictive models. These models demonstrated proficiency in identifying instances of infant sepsis prior to clinical recognition. Six of the models yielded a mean AUC between 0.800 and 0.820 for culture-positive cases, and for the cases identified either through culture or clinical diagnosis, the AUC ranged from 0.85 to 0.87. Data registry encompasses a myriad of variables such as demographics, vital signs, diagnoses, antibiotic usage, microbiological data, and treatment histories. Based on this registry, 618 unique infants were identified who collectively underwent 1188 sepsis evaluations, meeting all inclusion and exclusion criteria. A total of 110 culture-positive and 265 clinically diagnosed cases were identified from these evaluations. For predictive features, the model incorporated 36 variables that included clinical evaluations and indicators of comorbid conditions, which were identified through an extensive literature review and consultation with domain experts. Mean imputation was applied to handle missing data, and continuous features were standardized. Feature selection was automated through mutual information measures between each feature and the sepsis outcome. The study evaluated a range of machine learning models including NB, KNN, Gaussian processes, RF, AdaBoost, XGBoost, LR with L2 regularization, and SVM with a radial basis function kernel. Performance metrics used were AUC, sensitivity, specificity, PPV, and NPV. In the dataset focusing solely on culture-positive cases (CPOnly), the SVM model outperformed others across all metrics. However, no single model was superior in the dataset that also included clinically diagnosed cases (CP+Clinical). LR exhibited robust performance, achieving the highest average AUC along with AdaBoost for the CPOnly dataset and closely paralleling the top-performing gradient boosting models for the CP+Clinical dataset. Additional feature importance analysis revealed that logistic regression was more resilient to sample variability and less prone to overfitting compared to other high-performing, non-linear models. However, the study acknowledged certain limitations including potential bias in model learning curves (excluding logistic regression), the possible introduction of bias due to mean imputation, and the challenge of transforming retrospective decision-support frameworks into clinically viable tools.

The models' applicability beyond the NICU setting without retraining was also highlighted as a constraint. In summary, the study underscores the promise of machine learning models, particularly logistic regression, in facilitating early sepsis detection in infants.

The research study [37] aims to evaluate the effectiveness of a machine-learning algorithm in predicting the onset of sepsis in a pediatric demographic. This study made use of de-identified EHR data spanning emergency and inpatient encounters from 2011 to 2016, provided retrospectively by the University of California's San Francisco Medical Center. The machine learning algorithm showcased a notable AUC of 0.916 at the time of sepsis onset and 0.718 four hours before the onset, indicating effective differentiation between pediatric sepsis cases and control groups. Employing cross-validation and pairwise t-tests, the algorithm surpassed the PELOD-2 and SIRS scores in predicting severe sepsis four hours prior to the appearance of clinical symptoms. The patient data included were limited to ages between 2 and 17, deliberately excluding encounters with patients under 2 or older than 17. The rationale behind this age-specific selection was the inherently immature adaptive and innate immune systems in infants below the age of two. Of the 9486 encounters that were omitted from the study, 101 (or 1.060%) met the definition of severe sepsis. The algorithm was trained using multiple patient variables including age, blood pressure, heart rate, temperature, respiration rate, and peripheral oxygen saturation. Compared to PELOD-2 and pediatric SIRS scores, the machine-learning model exhibited superior performance, evidenced by higher AUC values and diagnostic odds ratios. Performance evaluation was carried out at multiple intervals, specifically one and four hours before PELOD-2 and zero, one, and four hours before SIRS, with hourly assessments commencing from the onset of sepsis to four hours prior. While the specific machine learning algorithm used for the study was not disclosed, its capability in effectively predicting the onset of severe sepsis in pediatric populations was statistically substantiated. By filling a research gap through the creation of an early warning system tailored for pediatric sepsis, this study constitutes a significant addition to the existing body of knowledge. The machine learning system offers continuous risk assessment by automatically monitoring EHRs, providing clinicians' valuable decision-making support that could potentially improve patient outcomes. Compared to traditional scoring systems, this machine-learning-based predictive model excelled in anticipating

severe pediatric sepsis. This could allow clinicians to identify at-risk pediatric patients earlier and allocate limited clinical resources more effectively to prevent adverse outcomes. The study suggests future research should focus on acquiring diverse and extensive datasets to refine the algorithm's performance and adapt it to each hospital's unique patient demographics. However, it is pertinent to note the study's limitations. Being retrospective and confined to a single tertiary care facility, its findings may not be universally applicable. Additionally, the use of ICD-9 codes as the gold standard could potentially miss some sepsis cases, and the algorithm's prediction was solely based on vital signs and laboratory data, excluding other clinically relevant information such as medical history or physical exam results. The study also lacks prospective validation, an aspect crucial for future research to determine the algorithm's clinical utility and effectiveness in enhancing patient outcomes.

The research paper under [38] investigates the utility of machine learning techniques in forecasting sepsis, a potentially lethal medical condition with diverse clinical manifestations that make its diagnosis and management complex. The authors critique traditional predictive systems for their inadequacy in swiftly identifying patient deterioration at an individual level. They advocate for machine learning algorithms capable of leveraging large and multifaceted healthcare datasets. The study reports that these algorithms outperform established scoring systems like SOFA and qSOFA, and were constructed using the MIMIC III database. The authors underscore the crucial role of judicious data, cohort, and feature selection in enhancing model accuracy. The authors propose the creation of a reference database and an ensemble classification technique rooted in machine learning to expedite the early identification of sepsis. The primary objectives of this paper are twofold: to pinpoint the most salient variables for early sepsis detection and to evaluate the relative performance of machine learning ensemble models against existing sepsis mortality scoring methods. To achieve these aims, the researchers crafted a systematic framework tailored for the early identification of sepsis in adult ICU patients. This framework is explicitly designed to efficiently detect organ dysfunction and suspected infection, aligning with the SEPSIS-3 criteria. The sepsis prediction model under scrutiny incorporates a broad array of pertinent factors, including vital signs, laboratory assessments, and structured demographic variables. Performance evaluation was undertaken within a one-hour predictive window, contrasting the outputs of diverse

machine learning algorithms against traditional scoring metrics. To maintain methodological rigor and consistency, the study included only subjects aged 14 or older, while purposefully excluding those admitted to the ICU with pre-existing sepsis to adhere to SEPSIS-3 guidelines. A total of 31 clinically relevant features were identified and subsequently categorized into three main domains: physiological data, laboratory findings, and demographic/score variables. Feature selection was executed through techniques such as information gain, relief, and Gini index. A range of machine learning algorithms—including SVM, KNN, ANN, NB, RF, AdaBoost, Stacking, and XGBoost—were applied to predict sepsis occurrences. Performance evaluation incorporated a spectrum of metrics such as AUC, precision, F1 score, recall, specificity, and overall accuracy. The results evinced that ensemble learning techniques displayed superior accuracy compared to standalone classifiers. Machine learning methods, particularly ensemble models coupled with feature selection procedures, were demonstrated to substantially enhance sepsis prediction precision. Additionally, the amalgamation of laboratory test results through cluster models yielded optimal performance. Specifically, within a 24-hour data window for one-hour prediction, the XGBoost model emerged as exceptionally efficacious, achieving an AUC value of 0.911. The authors suggest that deploying these machine learning models in ICU settings could significantly improve the timely detection of sepsis.

The objective of the research [39] is to develop an AI-based algorithm designed to facilitate early sepsis prediction, a complex problem faced by healthcare providers and systems globally. Utilizing a secondary analysis methodology, the study examined data from 4449 patients afflicted with infections and confined to the ICU at Zhengzhou University. The research employed a random forest algorithm incorporating a feature set of 55 variables extracted from electronic medical records. The predictive model exhibited promising results, boasting an AUC of 0.910, alongside a sensitivity of 87% and a specificity of 89%. Despite these robust outcomes, the authors advocate for independent validation studies to confirm the algorithm's adaptability across diverse populations and healthcare settings. The study was retrospective in nature and focused on data collected at the Zhengzhou University ICU from 2014 to 2016. The patient population was restricted to individuals 18 years and older, who met globally recognized criteria for sepsis and septic shock and were suffering from infection-related illnesses. Exclusion

criteria encompassed individuals younger than 18, diseases not related to infections, and incomplete data sets. Comprehensive infection and sepsis-specific medical and laboratory data were amassed for analysis. Statistical evaluations in the study included the use of counts and percentages for binary variables, assessed through Chi-square and Fisher's exact tests. Continuous variables were compared using Mann-Whitney and t-tests, while non-normal distributions were evaluated with the U-test. Results were presented as means accompanied by the standard error of the mean (SEM). The algorithm considered a total of 55 features related to various physiological and laboratory indicators such as lipid profiles, liver functionality, hem agglutination properties, and electrolyte levels. The model was fine-tuned through Gini importance and underwent training on different subsets of the available data, thereby enhancing its accuracy and robustness. The model demonstrated impressive discriminatory capabilities with an AUC score of 0.910. The study also accentuated the importance of electrolytes, specifically calcium, in predicting sepsis in infection-affected patients. However, the research acknowledges its own limitations, which include its retrospective design and a consequent vulnerability to information bias. It also admits to the lack of generalizability, given its focus on a Chinese patient population and the omission of other potential predictor variables. The authors suggest that further studies should employ a prospective design, include a broader and more diverse patient population, and incorporate additional predictor variables to enhance the algorithm's predictive accuracy and applicability. Further research is advocated, and the authors demonstrate scientific integrity by acknowledging the study's constraints, suggesting that future endeavors should seek to augment both the sample size and variable scope to improve predictive precision for sepsis events.

The study [40] explores the feasibility of applying machine learning algorithms to enhance the precision of predicting 30-day in-hospital mortality (IHM) among individuals presenting with suspected sepsis in the Emergency Department (ED). This approach outperforms traditional risk assessment tools, including qSOFA, NEWS, and MEWS. The study employed a secondary analysis approach, scrutinizing electronic health records of patients aged 21 or older, who were treated for suspected sepsis at Singapore General Hospital between September 2014 and April 2016. For inclusion, patients had to fulfill at least two out of the four SIRS criteria, which include temperature, heart rate, respiratory rate, and total white blood cell count. Data for the analysis was sourced from initial triage

records, including vital signs and ECG tracings. A range of machine learning methodologies such as SVM, AdaBoost, RF, and GBoost were employed in the study. The performance of these predictive models was evaluated and calibrated using metrics like the F1 score and the AUPRC. The research highlights substantial advancements in the prediction of 30-day IHM through the application of machine learning models, with a particular emphasis on gradient boosting techniques. An F1 score of 0.500 and an AUPRC of 0.350 stand as empirical proof of the superiority of these advanced models over traditional risk stratification instruments. However, the study is not without its limitations. These encompass its single-center design, limited sample size, exclusion of patients failing to meet at least two SIRS criteria, employment of a limited set of predictor variables, omission of specific interventions, and absence of an independent validation cohort. These factors collectively may affect the study's generalizability, statistical power, predictive accuracy, interpretability, and real-world applicability. As a remedial measure, future research would benefit from employing a more comprehensive, multi-center approach, featuring larger sample sizes and expanded inclusion criteria to enhance generalizability and robustness of the findings.

The research [41] explores the employment of machine learning-based models for forecasting IHM among individuals suffering from sepsis in ICUs. After scrutinizing five distinct algorithms, the study concluded that the GBDT model is the most efficacious for estimating mortality rates in ICU-admitted sepsis patients. The paper assesses five different algorithms (GBDT, LR, KNN, RF, and SVM) utilizing data from the MIMIC III dataset. In the subset of ICU patients afflicted with sepsis, the GBDT algorithm exhibited an exceptional capability for mortality prediction. Key performance indicators such as the AUC, recall, precision, and F1 score substantiate the model's superiority over alternative algorithms. Specifically, the AUC stood at 0.992, the recall rate was 94.8%, accuracy measured at 95.400%, and the F1 score was 0.933. Moreover, the performance of RF, SVM, and KNN models surpassed that of the LR model. The outcome of the GBDT model implies its potential utility in the development of future clinical decision support systems.

The article [42] delineates the advancements facilitated by data-driven methods, predominantly machine learning, in the diagnosis, subtyping, prognosis, and tailored treatments for sepsis. By focusing on diagnostic markers and electronic records, the article sheds light on how these data-driven techniques can improve assessment, clarify physiological pathways, streamline clinical trial recruitment, and refine clinical care strategies. Supervised and unsupervised methodologies show promise in pinpointing biomarkers linked to sepsis or its subtypes, which can in turn bolster diagnostic processes and suggest possible therapeutic interventions. There is a plethora of sepsis prediction methodologies in the current literature that leverage clinical data to pinpoint patients at risk of sepsis and septic shock. For instance, Mao et al. unveiled the InSight prediction tool which achieved impressive AUC values. A comprehensive meta-analysis from 2020 showcased the diverse range of AUC values across 130 models from various clinical contexts. Models such as COMPOSER, which relies on 40 clinical metrics, have shown robust performance in both ICU and ED settings. Furthermore, a substantial prospective multicenter study underscored the potential of real-time algorithms that consider a gamut of patient data, demonstrating their capability in early sepsis identification. The potential of biomarkers from transcriptomic, proteomic, and metabolomic studies in risk categorization and mortality predictions for sepsis patients is another salient aspect of the article. Transcriptomic data, for instance, has been utilized effectively in neural network algorithms to discern infections originating from bacteria and viruses. Inflammatory biomarkers have been consistently associated with sepsis' presence, progression, and outcomes. Moreover, metabolomics offers keen insights into the metabolic deviations seen in sepsis, revealing consistent disruptions in certain metabolic pathways. By harnessing gene-expression data, machine learning has also facilitated the categorization of sepsis patients based on diverse clinical outcomes and responses to treatments. Cluster analysis has revealed sepsis sub-phenotypes, offering novel insights into the pathology of sepsis and the potential treatment pathways. An intriguing dimension is the recognition of inflammatory sub-phenotypes through the monitoring of inflammatory mediators, as seen in ARDS, a recurrent sepsis complication. The article also discusses the overlap seen in COVID-19 induced latent classes and ARDS inflammatory phenotypes. In conclusion, the article accentuates the potential of machine learning in refining the diagnosis and management of sepsis. By harnessing vital signs, conventional clinical, and laboratory data, machine learning offers promise in early sepsis identification, therapeutic optimization, and targeted clinical trial enrollment. Nonetheless, challenges persist, such

as discrepancies in sepsis definitions and potential biases in training datasets. While machine learning promises enhanced precision in diagnosis and treatment, many existing models necessitate further validation. The majority of studies remain in preliminary stages, necessitating comprehensive research to verify their prospective clinical advantages.

The literature review [43] advocates for the utilization of machine learning, deep neural networks, and the updated SEPSIS-3 criteria to enhance predictive accuracy. A thorough analysis of 21 machine learning methodologies geared toward sepsis prediction unveiled disparities in definitions of sepsis, data source variations, preprocessing methods, differing strategies, feature engineering, and inclusion criteria. Notably, greater AUC results were observed closer to the onset of sepsis, largely attributed to the deployment of machine learning primarily for feature engineering. Employing DNNs combined with the SEPSIS-3 diagnostic criteria yielded superior results. The overarching aim is to refine machine learning models for early detection and intervention in sepsis cases. The review underscores the existing gaps in conventional prediction techniques, emphasizes the lack of consensus on datasets and feature processing methods, and accentuates the need for standardized benchmarks in the realm of medical research, especially for machine learning-based models. The literature review encompasses 21 pertinent studies: two are geared towards severe sepsis, while the remainder focus on early sepsis detection, its prediction, and mortality rate. Thirteen studies delved into preprocessing methodologies, with seven employing the MIMIC database. A mere six adopted the SEPSIS-3 definitions, with others opting for preceding standards. Remarkably, there was an absence of studies providing sample size rationales or methodological justifications. The quality of these studies was gauged using the JBI tool, revealing that a majority of machine learning models surpassed traditional prediction tools, with several achieving commendable AUC scores. The study underscores the application of machine learning in sepsis prediction, especially with a lens on mortality prediction and its time-sensitive nature. A common focus among most studies was the initial 24 hours following sepsis onset. Eight studies harnessed convolutional neural networks and ensemble learning algorithms, noting enhanced AUC outcomes as sepsis onset neared. The paper also delves into feature engineering methodologies used across 10 studies, examining clinical scores, demographic data, lab outcomes, and vital stats. The studies were bifurcated based on

their feature selection methods: either through specialized engineering techniques or through expert-led determinations. The adoption of diagnostic criteria in sepsis prediction studies was a salient theme of the review. Earlier definitions yielded high AUC outcomes but were found lacking in specificity and sensitivity due to their broader scope. Machine learning-based feature engineering emerged as a pivotal tool in identifying key factors for predictive models. Emphasis was placed on the importance of feature engineering in enhancing computational effectiveness, eliminating redundant data, and bolstering robustness. Despite the successes of domain-specific and machine learning-driven feature engineering in predicting sepsis-related mortality, there remain discrepancies in findings, necessitating further exploration. Studies predominantly leaned on feature extraction and specific modeling techniques for predicting sepsis mortality. While a few studies opted for clinical expertise to dictate features, this approach was potentially fraught with subjectivity and potential oversights. Others augmented their models using advanced techniques such as PCA, auto encoder-decoder structures, DFN, and LSTM. This comprehensive review underscores the potential and challenges of employing machine learning techniques for sepsis prediction. Highlighting the importance of addressing missing data, the study showcases the efficacy of neural network algorithms in early sepsis detection and mortality predictions. It offers insights into areas warranting further exploration, underscoring the transformative potential of machine learning. The heterogeneity of data in sepsis prediction models is examined, elucidating the challenges in discerning between afflicted and non-afflicted patients. The combination of clinical expertise with machine learning can yield high-frequency clinical data, simulating sepsis progression. The authors propose an advanced set of standards encompassing normalized datasets, feature engineering, evaluation criteria, and forward validation. Notably, none of the reassessed 21 models garnered a top rating, prompting a call for an internationally recognized benchmark for machine learning-centric models in clinical contexts.

The literature review [44], the modeling and statistical techniques adopted by machine learning in predicting sepsis among adult ICU patients are thoroughly assessed. Covering 14 comprehensive investigations, this review discerned various methodologies concerning sepsis definition, event determination, modeling parameters, and strategies. The prediction models showcased a diverse AUC spectrum, ranging from 0.610 to 0.960. Two notable studies emphasized that incorporating machine learning models can

potentially augment patient outcomes. However, there's an underscored need for further studies to set and evaluate standards for clinical method integration, emphasizing the role nurses can play in algorithm development. Sepsis remains a formidable health challenge, often leading to grave organ malfunction and fatalities, asserting its position as a predominant cause of in-hospital deaths. Despite advances in sepsis management, mortality rates have not seen a significant dip. The article highlights the evolution of sepsis definitions from the foundational SIRS criteria to the advanced SEPSIS-3 definitions. Machine learning's efficacy in rapidly detecting sepsis is established, but the adoption and protocols for its application in sepsis contexts are still under debate. The encompassed review concentrated on prospective studies targeting patients aged 15 and above from diverse medical settings, with an outcome focus on hospital-induced sepsis or shock. It considered articles from inception until October 1, 2018, with exclusions applied to non-English publications and specific study types. The search, centered on sepsis machine learning, tapped into databases like PubMed, CINAHL, and Cochrane Database of Systematic Reviews. Out of 465 identified publications, only 28 were deemed appropriate, with 14 being selected post a rigorous full-text review. These 14, derived from 23 articles, were US based, spanning 2010 to 2018, and predominantly included retrospective cohort studies. The variance in sepsis definitions, parameters, and modeling techniques was evident, with model AUCs oscillating between 0.610 to 0.960. An intriguing observation was the inconsistency in sepsis definitions and diagnostic timelines across the reviewed models. While some studies utilized SEPSIS-3 or SEPSIS-2 criteria, others leaned towards predicting septic shock or other parameters like a rise in SOFA score or SIRS criteria. In terms of model efficacy, the AUCs, despite different calculation techniques, displayed a range between 0.610 to 0.960. The review emphasized caution in solely relying on these figures.

An exhaustive overview of the academic publications central to the scope of the literature review is provided in Appendix 1.

3. METHODS

3.1. MIMIC III Database

An invaluable tool for research in the field of intensive care is the Medical Information Mart for Intensive Care III (MIMIC III) database. It provides a special compilation of anonymized health-related data spanning more than a decade, enabling multivariate analysis that could improve clinical decision-making procedures. Granular patient data from a variety of patient populations are stored in the MIMIC III database, an openly accessible critical care database. This enables thorough studies on a range of medical and healthcare research questions and includes demographics, vital signs, lab tests, medications, and more. Over 62135 critical care patients' data from the Beth Israel Deaconess Medical Center were collected for the MIMIC III database between 2001 and 2012 [45]. High temporal-resolution data from care provider notes, laboratory results, imaging reports, and electronic monitoring of vital sign measurements are all combined in this database, which covers admissions to intensive care units [46, 47].

A wide range of research fields have advanced as a result of the abundant repository of diverse, multidimensional health-related data in the MIMIC III database. Applications include everything from mortality prediction models and clinical decision support systems to finer analyses like drug response modeling and sepsis detection. Its use in the creation of machine learning models for the early sepsis prediction, has shown promising outcomes. For researchers researching critical care medicine and related fields, the MIMIC III critical care database is a useful tool. The database includes a variety of information that can be used to address significant research questions, such as patient demographics, laboratory test results, and clinical notes. The MIMIC III database has a lot of potential, but it also has some problems such as the ability to generalize results from MIMIC III may also be constrained because the dataset only includes data from one center.

The following tables are part of the MIMIC III database:

ADMISSIONS: Contains information about the patients stay, including the admission and discharge dates.

PATIENTS: Contains demographic data, including gender and birth date.

ICUSTAYS: Contains details about each ICU stay, including the time of admission and discharge.

SERVICES: Gives information about the service that a patient was accepted under.

TRANSFERS: Includes details on transfers to and from various wards / units.

DRGCODES: Includes details on Diagnosis Related Group (DRG) codes.

DIAGNOSES_ICD: Contains patient diagnoses listed according to the International Classification of Diseases (ICD).

PROCEDURES_ICD: Contains procedures that have been ICD-coded.

PRESCRIPTIONS: Includes all prescriptions issued by hospitals.

LABEVENTS: Contains patient-specific laboratory measurements.

OUTPUTEVENTS: Contains patient outputs (e.g., from a urine output).

INPUTEVENTS_CV and *INPUTEVENTS_MV*: Contain records of the fluid intake activities experienced by patients, split into two separate tables due to variations in data collection.

DATETIMEEVENTS: Contains all occurrences with dates and times noted, such as when a patient leaves their bed.

CHARTEVENTS: Contains all patient observations that have been charted.

MICROBIOLOGYEVENTS: Includes tests for microorganisms and the corresponding sensitivities.

NOTEEVENTS: Contains notes that have been de-identified, such as notes from nurses and doctors, ECGs, imaging reports, and summaries of discharges.

CAREGIVERS: Contains details about the caregivers who entered information into the database.

CALLOUT: Contains details regarding the patients' readiness for discharge.

CPTEVENTS: Includes procedures coded in the Current Procedural Terminology format (CPT).

D_CPT: Contains extensive details regarding every Current Procedural Terminology (CPT) code.

D_ICD_DIAGNOSES and *D_ICD_PROCEDURES*: Contain high-level details about each ICD code.

D_ITEMS and *D_LABITEMS*: contain in-depth knowledge about objects and lab equipment.

The following MIMIC III database tables were used in the context of the study.

Table 1. ICUSTAYS

ICUSTAYS	
Field	**Description**
subject_id	Unique identifier of the patient
hadm_id	Hospital admission number of the patient
icustay_id	ICU stay ID
intime	Admission time of the patient to the ICU
outtime	Time of discharge from the ICU

Table 2. LABEVENTS

LABEVENTS	
Field	**Description**
subject_id	Unique identifier of the patient
hadm_id	Hospital admission number of the patient
charttime	The time when the laboratory test was performed
itemid	Unique identifier of the test.
value	The original value of the test result
valuenum	Numerical value of the test result

Table 3. CHARTEVENTS

CHARTEVENTS	
Field	**Description**
subject_id	Unique identifier of the patient
hadm_id	Hospital admission number of the patient
icustay_id	ICU stay ID
charttime	The time when the event (chart) occurre
itemid	Identifier for a single item that was charted
valuenum	The numerical results of the item that was charted.
error	Indicates whether the chart event was marked as an error or not

Although it acknowledges the potential advantages of hospital data in enhancing patient care, problems with data integration and digital system interoperability have prevented its full realization. An hourly data sets were produced within the parameters of the study with the aid of the subsequent queries.

To compile arterial blood gas information from the sepsis study that is relevant to ICU stays, a database table with the name arterialbg icustays has been created. It is based on various *ITEMID*s that purport to correlate with various lab measurements. Additionally, these *ITEMID*s have labels assigned to them for improved identification. To ensure that they are in line with expected ranges, the values in question have undergone thorough sanity checks. A number of indexes have been generated for this table since it was created. The subject *id, hadm id, icustay id, charttime, intime, outtime, sepsis3,* and *label* indexes are designed to speed up searches through this table.

3.1.1. Designing the Comprehensive Dataset for Study Analysis

In the context of this research, the dataset under scrutiny was meticulously constructed through the execution of specific queries within the MIMIC III database, a publicly accessible critical care database.

Steps of the query are as followings:

1. **Table Creation** ("*sepsis.arterialbg_icustays*"): The "*icustays*" and "*labevents*" tables, along with the "*public.sepsis3*" table, are being combined to create a new table called "*arterialbg icustays*" in the "*sepsis*" schema.

"*sepsis3*" table: This table contains information on whether there is sepsis according to SEPSIS-3 criteria.

Table 4. sepsis3

sepsis3	
Field	**Description**
hadm_id	Hospital admission number of the patient.
icustay_id	ICU stay ID
suspected_infection_time_poe_days	Duration when an infection was suspected in the patient
sofa	Sequential Organ Failure Assessment (SOFA) score
age	Patient age
outtime	The time of departure from the ICU
intime	The time of admission to the ICU
suspected_infection_time_poe	The time of suspected infection
specimen_poe	The specimen taken from the patient
positiveculture_poe	Whether a positive culture is present
antibiotic_time_poe	The start time of antibiotics
blood_culture_time	The time when the blood culture was taken
blood_culture_positive	Whether the blood culture is positive.
gender	The gender of the patient
diabetes	Whether the patient has diabetes

first_service	The first service provided to the patient
hospital_expire_flag	Death upon hospital discharge
thirtyday_expire_flag	Death within 30 days.
icu_los	The duration of stay in the ICU
hosp_los	The duration of hospital stay
sofa, lods, sirs, qsofa, qsofa_sysbp_score, qsofa_gcs_score, qsofa_resprate_score	Various health scores

a. **Data Cleaning:** The *valuenum* column is subject to certain checks, and some values are assigned as NULL. For instance, observations with these values are labeled as NULL when they have negative values, hematocrit values greater than 100, FiO2 (Fraction of Inspired Oxygen) values greater than 20 or less than 100, O2 saturation values greater than 100, O2 flow values greater than 70, or PO2 values greater than 800.

b. **Data Labeling**: The names of the lab tests are assigned to the label column based on the values in the *itemid* column.

c. **Time Interval Calculation:** The *charttime_intime_intv* column is calculated as the interval between the *intime* (time of ICU admission) and *charttime* (time of lab test) columns.

d. **Sepsis Status Calculation:** Based on the suspected infection time POE (Physician Order Entry) days and sofa columns, the Sepsis3 status is determined. Sepsis3 status is considered to be 1, if suspected infection time POE days is not NULL and sofa value is 2 or higher, otherwise it is considered to be 0.

2. **Join Operation:** Based on matching "*subject id*", "*hadm id*", and "*icustay id*" values as well as the requirement that the "*charttime*" in "*labevents*" falls between the "*intime*" and "*outtime*" in "*icustays*”, the "*icustays*" table is joined with the "*sepsis3*" and "*labevents*" tables.

3. **Selection and Transformation:** The "*itemid*" column is converted to human-readable labels using a case statement, and a new column named "*charttime intime intv*" is created as the distinction between "*charttime*" and "*intime*".

Table 5. sepsis.arterialbg_icustays

sepsis.arterialbg_icustays	
Field	**Description**
icustay_id	ICU stay ID
hadm_id	Hospital admission number of the patient
subject_id	Unique identifier of the patient
age	Patient age
charttime	The time when the event (chart) occurre
charttime_intime_intv	The interval between the intime (time of ICU admission) and charttime (time of lab test) columns
intime	The time of admission to the ICU
outtime	The time of departure from the ICU
sepsis3	Sepsis3 status is considered to be 1, if suspected infection time poe days is not NULL and sofa value is 2 or higher, otherwise it is considered to be 0.
feature	This is taken from the "label" column and some feature names have been renamed (TEMPERATURE → Temp, SO2 → SpO2, GLUCOSE → Glucose).
valuenum	Numerical value of the test result.

The sepsis schema has generated a table with the name labs_icustays. This table contains thorough information about laboratory activities that took place while patients were being treated in the ICU for suspected sepsis. The information contained in this table was taken from the MIMIC III database's *icustays*, *sepsis3*, and *labevents* tables.

The results of numerous laboratory tests are included in the "*labs icustays*" table. As indicated by their item IDs, the query specifically curate's data related to a predetermined range of tests, including Anion Gap, Albumin, Bicarbonate, and Bilirubin, among others. These tests are frequently used in the identification or management of sepsis.

In essence, the creation of the "*labs icustays*" table was done in order to make it simple to compare lab results with ICU stays and sepsis cases. It creates a table with relevant lab data that is properly labelled, filtered, and optimized for quick querying. This could be useful in a variety of analyses, such as examining the connections between the results of various laboratory tests and the prevalence of sepsis.

Steps of the query as followings:

1. Input data: It retrieves information from the database's "*icustays*," "*labevents*", and "*sepsis3*" tables.
2. Joining Tables: Based on the "*hadm id*" and "*icustay id*" identifiers, joins are created between the "icustays" and "sepsis3" tables. In addition, it joins the "labevents" table using the "*subject id*", "*hadm id*", and "*charttime*" identifiers as well as the times of the lab events during the ICU admission period ("*intime*" and "*outtime*").
3. Filtering data: The script also checks the value of these tests ("*valuenum*") is not null and greater than 0 and filters "*labevents*" for a specific set of *ITEMID*s that represent particular lab tests.
4. Transforming data: The script carries out the following transformation tasks:
    a. It substitutes a human-readable label designating the kind of lab test for the *ITEMID*.
    b. It verifies the accuracy of the lab findings and sets any irrational values to zero.
    c. It determines a variable called "*charttime intime intv*," which denotes the amount of time between ICU admission and the lab event.

d. On the basis of the presence of a suspected infection and the SOFA score, a binary variable called "*sepsis3*" is added.

5. Creating a new table: The script loads all of the gathered and transformed data into a new table called "sepsis.labs icustays".

Table 6.sepsis.labs_icustays

sepsis.labs_icustays	
Field	**Description**
icustay_id	ICU stay ID
hadm_id	Hospital admission number of the patient
subject_id	Unique identifier of the patient
age	Patient age
charttime	The time when the event (chart) occurre
charttime_intime_intv	The interval between the intime (time of ICU admission) and charttime (time of lab test) columns
intime	The time of admission to the ICU
outtime	The time of departure from the ICU
sepsis3	Sepsis3 status is considered to be 1, if suspected infection time poe days is not NULL and sofa value is 2 or higher, otherwise it is considered to be 0
feature	This is taken from the label column and some feature names have been renamed (GLUCOSE --> Glucose)
val	This is taken from the "valuenum" column

To create a new table ("*vitals icustays*") for the sepsis schema, another query has been created. This table is derived from the MIMIC III database's "*icustays*" and "*chartevents*" tables as well as the "*sepsis3*" table from the public schema. The "*vitals icustays*" table

will contain a variety of vital sign measurements related to each ICU stay, including heart rate, blood pressure, respiration rate, oxygen saturation (SpO2), glucose level, and body temperature [48]. The "*chartevents*" table contains a number of item IDs for each vital sign, each of which corresponds to a different way of measuring or recording that particular vital sign.

Steps of the query as followings:

1. Input data: It retrieves information from the database's "*icustays*", "*chartevents*", and "*sepsis3*" tables.
2. Join Operation: There are three joined tables: "*icustays*", "*chartevents*", and "*sepsis3*" tables.
3. The main join criteria are "*subject id*", "*hadm id*", and "*icustay id*."
4. Filtering: Rows are filtered according to specific criteria, such as whether "*ce.charttime*" falls between "*ie.intime*" and "*ie.outtime*", and specific item ID standards.
5. Transformation and Calculation: The conversion of temperatures from Fahrenheit to Celsius, the calculation of the charttime-in-time interval, and the creation of "*VitalID*" and "*VitalName*" based on item ID are a few examples of fields that have been transformed or calculated based on the data already present.
6. Table Creation: The finished item is saved as a brand-new table in the "*sepsis*" schema with the name "*vitals icustays*".

Table 7. sepsis.vitals_icustays

sepsis.vitals_icustays	
Field	**Description**
icustay_id	ICU stay ID
hadm_id	Hospital admission number of the patient

subject_id	Unique identifier of the patient
age	Patient age
charttime	The time when the event (chart) occurre
charttime_intime_intv	The interval between the intime (time of ICU admission) and charttime (time of lab test) columns
intime	The time of admission to the ICU
outtime	The time of departure from the ICU
sepsis3	Sepsis3 status is considered to be 1, if suspected infection time poe days is not NULL and sofa value is 2 or higher, otherwise it is considered to be 0
feature	This is taken from the vitalname column
val	This is taken from the valuenum column

A subsequent structured query creates the comprehensive table known as "*all icustays*" within the "*sepsis*" schema. This table combines information from three separate tables called "*vitals icustays*", "*arterialbg icustays*", and "*labs icustays*" each of which represents a different category of medical data gathered during ICU stays. The "*all icustays*" table serves as a compendium of diverse medical information gathered over the course of ICU stays. This table's entries each include a feature group label that identifies the type of data it contains, such as Vital Sign, Arterial Blood Gas, or Lab.

Steps of the query as followings:

1. Union Operation: Each table must have the same number of columns and compatible types for those columns in order to use this operation to combine the rows from multiple tables into a single table. Rows from "*vitals icustays*," "*arterialbg icustays*," and "*labs icustays*" are combined in this query.
2. Selection and Transformation: Specific columns are chosen for each table, some fields are renamed (for example, "*v.vitalname*" becomes "*feature*"), and new

fields are created based on the data already present (for example, "*Vital sign*", "*Arterial Blood Gas*", and "*Lab*" become "*feature group*").

3. Filtering: Certain requirements must be met in order to filter rows, such as that "*v.vitalid*" and "*a.label*" cannot be null.

4. Table Creation: The final outcome of the union and transformations is saved as a new table in the "*sepsis*" schema called "*all icustays*".

Table 8. sepsis.all_icustays

sepsis.all_icustays	
Field	**Description**
icustay_id	ICU stay ID
charttime	The time when the measurement was taken
charttime_intime_intv	The time interval from the initial ICU admission to the time of measurement
feature	The feature measured (for example, heart rate, respiratory rate, temperature, blood pressure, etc.)
val	The value of the feature

With a focus on patients with sepsis, the query that creates the new table "*icustay hourly features*" in the database's "*sepsis*" schema aims to condense the clinical parameters and features of ICU stays at an hourly interval. These parameters and traits are derived from the main table "*sepsis3*", and the hourly measurements are made easier by a second table called "*sepsis.all icustays*". The "*icustay hourly features*" query's main goal is to create a dataset specifically suited for analyses or model development targeting sepsis patients.

The MIMIC III database offers an exhaustive suite of individual patient records, chronicling each phase from admission to discharge in ICUs. Updated every hour, this data is finely segmented and can be efficiently harvested using query mechanisms. Such temporal granularity is particularly indispensable for diagnosing and managing rapidly

evolving conditions like sepsis. By continually assessing this high-resolution data, healthcare providers can discern critical variations in patient conditions, thereby fine-tuning interventions and treatment modalities in real time.

Time-sensitive management is pivotal in life-threatening conditions like sepsis, which demand rapid diagnosis and immediate therapeutic measures such as antibiotics and fluid resuscitation. Each elapsed hour without appropriate intervention aggravates the risk of progressive organ failure and subsequent mortality. Thus, in ICUs, it is imperative for clinicians to expedite decisions using real-time data and persistent monitoring, while also adapting treatments based on continual assessment of the patient's evolving condition.

The study under discussion leveraged the MIMIC III database to closely inspect the first 24 hours of each ICU admission, dividing the period into discrete hourly datasets. This enables healthcare teams to perceive subtle but significant shifts in patient states, facilitating more precise treatment planning. Regular, hourly assessments allow for agile adjustments in treatment, aligning closely with each patient's specific needs at any given hour. Ultimately, this temporally granular approach augments the odds of successful early diagnosis and treatment, thereby optimizing patient outcomes.

3.2. Sepsis Scoring Systems

Sepsis scoring frameworks are instrumental in medicine, furnishing clinicians with essential instruments to gauge the likelihood of sepsis onset and ascertain its severity post-manifestation. Integrating various clinical parameters, including symptomatic presentations, vital metrics, and laboratory test results, these frameworks afford a uniform and methodical methodology for evaluating and risk-stratifying patients concerning sepsis [49]. This facilitation of early detection and intervention is pivotal for efficacious sepsis management. Numerous sepsis scoring algorithms have been formulated and are routinely employed in clinical settings, each characterized by distinct merits and shortcomings. These models are subject to incessant refinement and validation, guided by emergent research and data, thereby ensuring their continued relevance across different patient demographics and healthcare environments. As invaluable adjuncts to

clinical decision-making, sepsis scoring systems assist physicians in astutely diagnosing sepsis, gauging its severity, and tailoring suitable treatment courses. The judicious deployment of these frameworks thereby enhances the quality of sepsis care, thereby optimizing patient outcomes [50].

3.2.1. Sequential Organ Failure Assessment (SOFA):

The SOFA score keeps tabs on how the body's various organ systems, such as the respiratory, cardiovascular, hepatic, coagulation, kidney, and central nervous systems, are doing. A total SOFA score is calculated by adding the scores for each organ system, which range from 0 (normal) to 4 (high degree of dysfunction/failure) [51]. A higher score denotes a higher risk of mortality and more severe organ dysfunction.

The SOFA scoring table is structured as follows:

Central Nervous System:

Evaluated by the Glasgow Coma Scale (GCS).

Scores range from 0 to 15, with lower scores indicating greater dysfunction [52].

Cardiovascular System:

Assessed by the Mean Arterial Pressure (MAP) or the requirement of vasopressor administration.

Respiratory System:

Assessed using the PaO2/FiO2 ratio, representing arterial oxygen partial pressure to fractional inspired oxygen ratio [53].

Coagulation:

Evaluated through the platelet count (×103/µl).

Liver:

Assessed based on the total bilirubin level (mg/dl) [µmol/L].

Renal Function:

Assessed by the creatinine level (mg/dl) [µmol/L] or urine output.

Table 9. SOFA

	Central nervous system	**Cardiovascular system**	**Respiratory system**	**Coagulation**	**Liver**	**Renal function**
Score	Glasgow coma scale	**Mean arterial pressure OR administration of vasopressors required**	**PaO_2/FiO_2 [mmHg (kPa)]**	**Platelets ($\times 10^3/\mu l$)**	**Bilirubin (mg/dl) [µmol/L]**	**Creatinine (mg/dl) [µmol/L] (or urine output)**
0	15	MAP ≥ 70 mmHg	≥ 400 (53.3)	≥ 150	< 1.2 [< 20]	< 1.2 [< 110]
1	13–14	MAP < 70 mmHg	< 400 (53.3)	< 150	1.2–1.9 [20-32]	1.2–1.9 [110-170]
2	10–12	dopamine ≤ 5 µg/kg/min or dobutamine (any dose)	< 300 (40)	< 100	2.0–5.9 [33-101]	2.0–3.4 [171-299]
3	6–9	dopamine > 5 µg/kg/min OR epinephrine ≤ 0.1 µg/kg/min OR norepinephrine ≤ 0.1 µg/kg/min	< 200 (26.7) **and** mechanically ventilated including CPAP	< 50	6.0–11.9 [102-204]	3.5–4.9 [300-440] (or < 500 ml/day)
4	< 6	dopamine > 15 µg/kg/min OR epinephrine > 0.1 µg/kg/min OR norepinephrine > 0.1 µg/kg/min	< 100 (13.3) **and** mechanically ventilated including CPAP	< 20	> 12.0 [> 204]	> 5.0 [> 440] (or < 200 ml/day)

3.2.2. Systemic Inflammatory Response Syndrome (SIRS)

The white blood cell count, respiratory rate, body temperature, and heart rate are all SIRS criteria. If a patient satisfies two or more of the requirements, they are said to have SIRS. Even though SIRS is not unique to sepsis and can be brought on by other inflammatory diseases, it is frequently used as a component of the assessment for sepsis.

Four criteria make up the SIRS scoring table, each of which corresponds to distinct clinical findings:

Temperature:

If the temperature is either 36°C (96.8°F) or >38°C (100.4°F), a score of 1 is given.

Heart Rate:

If the heart rate is greater than 90 beats per minute, a score of 1 is awarded.

Respiratory Rate:

If the respiratory rate is greater than 20 breaths per minute or the partial pressure of carbon dioxide (PaCO2) is lower than 32 mmHg, a score of 1 is given (4.3 kPa) [54].

White Blood Cell Count (WBC):

Based on the WBC count, various score values are assigned:

If the WBC count is less than $4x10^9$/L (4000/mm^3), it receives a score of 1.

If the WBC count is higher than 12x109/L (>12,000/mm3), a score of 1 is also given.

If there is an increase in immature white blood cells (band cells) equal to or greater than 10% of the overall white blood cell count, a second score of 2 is given.

Table 10. SIRS

Systemic inflammatory response syndrome	
Finding	**Value**
Temperature	<36 °C (96.8 °F) or >38 °C (100.4 °F)
Heart rate	>90/min
Respiratory rate	>20/min or PaCO2<32 mmHg (4.3 kPa)
WBC	<4x109/L (<4000/mm3), >12x109/L (>12,000/mm3), or ≥10% bands

3.2.3. qSOFA (Quick SOFA)

A simplified version of the SOFA score called the qSOFA score was developed for quick bedside evaluation. It meets three requirements: a changed state of consciousness, rapid breathing (>= 22 breaths per minute), and low blood pressure (systolic =100 mmHg). In the case of a suspected infection, a patient is deemed to be at high risk for a poor outcome if they meet two or more of these criteria.

Table 11. qSOFA

Assessment	**qSOFA score**
Low blood pressure (SBP ≤ 100 mmHg)	1
High respiratory rate (≥ 22 breaths/min)	1
Altered mentation (GCS ≤ 14)	1

3.2.4. National Early Warning Score (NEWS)

The NEWS serves as a risk-assessment tool designed for flagging individuals who may experience medical decline, inclusive of those susceptible to sepsis, despite its lack of sepsis specificity. This system appraises six key physiological parameters: respiratory rate, oxygen saturation levels, core body temperature, systolic blood pressure, heart rate, and level of consciousness as measured by the AVPU (Alert, Verbal, Pain, Unresponsive) scale. There is a statistically supported relationship indicating that a NEWS score of 5 or above is linked to an elevated likelihood of either mortality or admittance to an intensive care unit.

Table 12. NEWS

Score	**3**	**2**	**1**	**0**	**1**	**2**	**3**
Respiratory rate (breaths/min)	>35	31–35	21–30	9–20			<7
SpO2 (%)	<85	85–89	90–92	>92			
Temperature (C)		>38.9	38–38.9	36–37.9	35–35.9	34–34.9	<34
Systolic BP (mmHg)		>199		100–199	80–99	70–79	<70
Heart rate (bpm)	>129	110–129	100–109	50–99	40–49	30–39	<30
AVPU				Alert	Verbal	Pain	Unresponsive

3.2.5. APACHE (Acute Physiology and Chronic Health Evaluation)

In the realm of critical care medicine, particularly for the diagnostic evaluation of sepsis cases, the APACHE scoring system is widely implemented. Aimed at patients in ICUs, this system quantifies the severity of a patient's condition to predict mortality risk [55]. The calculation of the APACHE score incorporates an array of physiological measurements—ranging from vital signs to lab results—as well as age and any enduring health conditions.

3.3. SEPSIS – 3

The third iteration of the diagnostic criteria for sepsis, commonly known as SEPSIS -3, was established by the Third International Consensus Definitions for Sepsis and Septic Shock committee in 2016 [56]. This update sought to enhance the precision of sepsis identification and furnished a more detailed criterion framework for both clinical application and scholarly inquiry.

In the SEPSIS-3 criteria, sepsis is delineated "*as a life-threatening*" organ malfunction emanating from a dysregulated host reaction to infection. The criterion employs the SOFA score to gauge dysfunction across six key organ systems: respiratory, cardiovascular, hepatic, coagulation, renal, and neurological. An uptick of 2 points or more from the baseline SOFA score serves as an indicator of sepsis. Additionally, sepsis may escalate to septic shock even post-adequate fluid resuscitation; this escalation is marked by sustained hypotension and lactate concentrations exceeding 2 mmol/L, which signals tissue hypoperfusion.

Utilizing the SEPSIS-3 framework enables clinicians to diagnose sepsis with greater precision, thereby facilitating earlier, targeted interventions for enhanced patient outcomes. The incorporation of the SOFA score in the SEPSIS -3 criteria accentuates the centrality of organ dysfunction in the assessment of sepsis. By focusing on this key aspect, the criteria aim to more proficiently identify patients requiring intensive care and who are at elevated risk for adverse outcomes.

In the study, particular emphasis was laid on the SOFA scoring mechanism as well as the contemporaneous SEPSIS -3 criteria. Both these elements were deemed indispensable for the accurate delineation and assessment of patients manifesting sepsis within the research framework. By focusing on organ dysfunction as a quintessential parameter, both the SOFA scoring system and the SEPSIS-3 definition aim to refine the diagnostic precision and objectivity with regard to sepsis.

3.4. Methodological Approach to Data Acquisition and Analysis

The diagnostic paradigm for sepsis within the confines of this research is predicated on the SOFA score, a gradated scale from 0 to 4 that quantifies the extent of organ failure. Data for this study was extracted from the MIMIC III database, an expansive archive of patient-related information, selected with meticulous query-based criteria. SOFA scores were calculated, and the pertinent medical records for each patient's initial 24-hour period in the ICU were aggregated, considering each hour as an isolated timeframe. This culminated in an exhaustive dataset encompassing laboratory outcomes, vital statistics, and demographic indicators for a total of 61532 ICU-admitted patients. Subsequent to this, the data was partitioned into three age-specific cohorts: an all-encompassing age cohort, an elder cohort, and an infant cohort. For each of these cohorts, the dataset features hourly patient data for the first 24 hours subsequent to ICU admission.

Within the ambit of the research, two analytical paradigms—connected and non-connected methodologies—were deployed for evaluating the 24 individual datasets, each of which corresponded to the first 24-hour time frame post ICU admission for the three patient subgroups. The juxtaposition aimed to scrutinize the ramifications of acknowledging temporal interdependencies in the evolution of a patient's health status.

In the connected approach, the patient's health condition is conceptualized as a temporally contiguous sequence, rather than an aggregation of discrete, hourly events. In this framework, the status of the ailment is not isolated to individual hourly data points but is perceived as a dynamic continuum that undergoes transformation over successive time intervals. This temporally-informed model accommodates the variations in the patient's medical parameters from preceding hours, thereby facilitating a more nuanced apprehension of disease trajectories and therapeutic responses.

Conversely, the non-connected approach posits each hourly dataset as an autonomous entity, eschewing any causal relationships or statistical correlations between adjacent time periods. In this setting, the state of the disease remains uninfluenced by preceding

medical conditions and the analytical procedures are executed in abstraction from the time-related dimensions of the data.

In contrasting the outcomes from both the connected and non-connected methodologies, the research aimed to discern whether the utilization of temporal dynamics within the connected framework results in more accurate and nuanced understandings relative to the non-connected paradigm. This evaluative task is indispensable for gauging the import of temporal dependencies, specifically in the realms of prognosticating disease course and fine-tuning healthcare protocols within an ICU context.

Notably, the connected model prioritizes what is termed the "*confidence level*", a metric indicating the likelihood of the patient's compromised health in preceding timeframes. The integration of this confidence level amplifies the depth and breadth of understanding one gains about the patient's health trajectory. Such temporal dependencies furnish invaluable insights into the longitudinal unfolding of the disease state. Consequently, a patient's initial health profile upon ICU admission becomes an integral component in calculating the aggregate likelihood of their disease status at any given future point. This aspect of the connected approach is of heightened relevance when dealing with time-sensitive illnesses like sepsis, necessitating immediate and effective medical interventions. The said confidence level serves as a significant gauge for the progression of such swiftly escalating conditions, thus wielding influence over medical decision-making and strategy formulation.

Conversely, the non-connected paradigm perceives each hourly dataset as a discrete, isolated entity, implying no causal or temporal relationship with the patient's health in antecedent or subsequent timeframes. This paradigm operates on the presumption of independence among these temporal data points, thereby excluding any potential influences of temporal patterns or sequential correlations in disease development. In doing so, it engages with each hourly data capture as an isolated event, neglectful of how the patient's health might have dynamically evolved in a temporal context.

While the non-connected model offers the advantages of computational simplicity and reduced computational burden, it runs the risk of neglecting essential data obtainable through the consideration of temporal dependencies. For conditions characterized by rapid deterioration and demanding immediate clinical response, such as sepsis, the non-connected framework might be insufficient in encapsulating the dynamic complexities of the disease. This limitation could conceivably compromise the accuracy of both prognostic evaluations and therapeutic decisions.

Conversely, the connected methodology, as elaborated upon earlier, integrates temporal dependencies into its analytical framework. It incorporates a metric known as the 'confidence level,' which reflects the patient's prior disease state, to influence the current understanding of their health condition. This inclusion enhances the model's capacity to apprehend the time-sensitive characteristics of certain diseases, thereby providing a more precise foundation for clinical decision-making.

3.5. Characteristics and Attributes of the Employed Dataset

In the study, two distinct analytical frameworks—termed the "*non-connected*" and "*connected*" models—were formulated for each of the trio of patient subgroups under study [10]. This resulted in an aggregate of 72 distinct datasets that were subsequently analyzed utilizing both the novel non-connected and connected model methodologies, as delineated in Table 13.

The utilization of hourly data configurations affords a robustly detailed temporal portrait of a patient's evolving medical condition, thereby enabling a more nuanced and dynamic appraisal of their health status. Such meticulous monitoring is instrumental in the early identification of conditions like sepsis, a critical factor in facilitating prompt medical intervention and thereby enhancing patient prognosis.

The employment of hourly data analysis techniques allows for the discernment of nuanced temporal patterns and trajectories, potentially signaling the emergence or

exacerbation of conditions such as sepsis. This elevated granularity in data assessment equips healthcare practitioners with invaluable perspectives into the dynamic morphology of the illness, thus aiding in the refinement of increasingly accurate predictive models.

On the flip side, utilizing an aggregated 24-hour data set could risk ignoring vital variations or shifts in the patient's health status that transpire on an hourly basis. This could lead to deferred identification of sepsis or a diminution in the precision of outcome forecasts.

The research study aimed to contrast and critically assess the relative capabilities and success of these two distinct modeling techniques in prognosticating both the health condition and eventual medical outcomes for patients within the different cohorts examined.

The data preparation process was carefully designed to ensure that both the non-connected and connected models could be applied to the respective data sets accurately and consistently.

Table 13 represents the distribution of data points across three distinct cohorts: Infant, Elder, and All Age. The Infant Cohort consists of 8100 data points, which are distributed across 24 individual data sets. The Elder Cohort comprises a substantially larger pool of data points, totaling 37069, again disseminated across 24 data sets. The All Age Cohort, which integrates data from both previous cohorts and possibly additional age groups, contains the largest volume of data points at 61532, equally divided across 24 data sets. In sum, there are 72 distinct data sets across all cohorts.

Table 13. The data sets consulted for the study

Name of Data Sets	**Number of Data Points**	**Number of Data Set**
Infant Cohort	8100	24

Elder Cohort	37069	24
All Age Cohort	61532	24
Total		72

The methods adopted are illustrated in Figure 1.

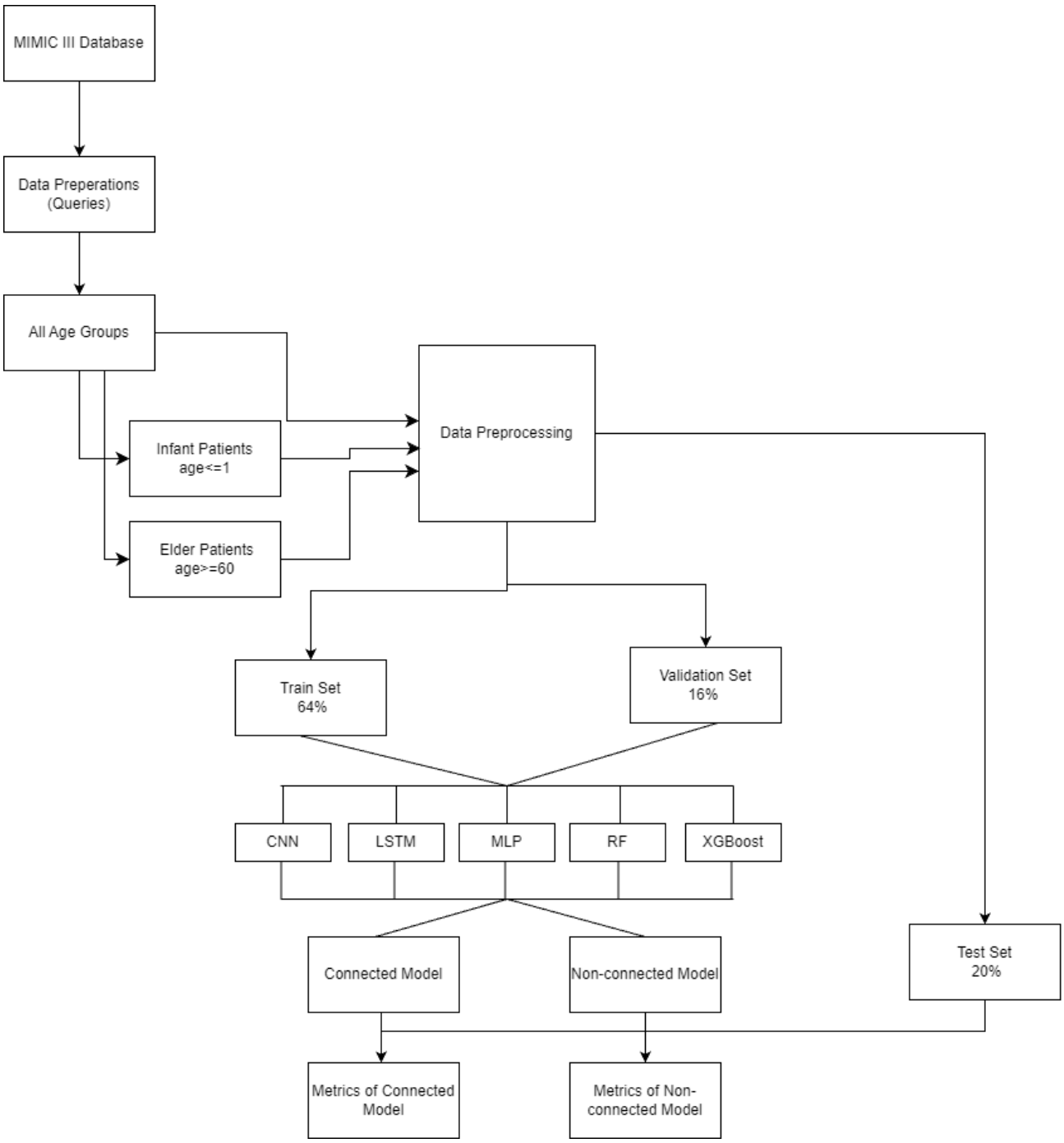

Figure 1. Methods

3.5.1. Data Sets Attributes

The complete data set comprises 18 variables, which were employed in the diagnostic process of sepsis among patients admitted to the ICU. The attributes of the data set are detailed in Table 14 [10].

Table 14. The attributes of the data sets

Metadata	Attribute	Type	Missing Percentage
icustay_id	Unique ID of Patients' stay at ICU	Admission	0%
sepsislabel	Label of Sepsis	Target	0%
age	Age	Demographic	0%
heartrate	Heart rate	Vital	39.5%
respiratory	Respiratory rate	Vital	49.9%
temp	Temperature	Vital	55.3%
sysbp	Systolic Blood Pressure	Vital	51.1%
spo2	Oxygen Saturation	Vital	49.5
diasbp	Diastolic Blood Pressure	Vital	51.1%
meanbp	Mean Arterial Blood Pressure	Vital	51.2%
wbc	White Blood Cell Count	Lab Results	69.3%
bun	Blood Urea Nitrogen	Lab Results	78.0%
creatinine	Creatinine	Lab Results	78.8%
ph	Arterial pH	Lab Results	72.0%
intime	Date and time of patients' entry in the ICU	Admission	0%
outtime	Date and time of patients' discharge from ICU	Admission	0%
suspected_ infection_ time_ poe	Time of Suspected Infection	Admission	33.3%

sofa	SOFA Score	Sequential Organ Failure Assessment Score	0%

Table 14 encapsulates metadata regarding various attributes, their type, and the respective percentage of missing data for each. Notably, the parameters that manifest the minimum missing percentage (0%) are "*icustay_id*", "*sepsislabel*", "*age*", "*intime*", "*outtime*", and "*sofa*". These attributes encompass the unique ID of a patient's stay in the ICU, the label of sepsis, demographic age, the date and time of a patient's entry and discharge from the ICU, and the SOFA score. These fields are completely recorded for every patient, and thus provide a comprehensive dataset for these specific variables.

On the other end of the spectrum, the parameters with the highest missing percentage are "*bun*" (Blood Urea Nitrogen) and "*creatinine*" (Creatinine), with missing data recorded at 78.0% and 78.8% respectively. Both of these attributes fall under the category of lab results, indicating that these specific tests were not consistently conducted, or the results were not consistently recorded, for all patients in the study.

The scrutiny of patient data on an hourly basis revealed the presence of incomplete data entries, most noticeably within the dataset encompassing all age groups, where a considerable volume of missing data was detected. To rectify this challenge, the technique of mean imputation was administered to substitute these absent values. This process involves the replacement of missing entries with the arithmetic average of the pertinent variable. Mean imputation was opted for given its merits, which include the preservation of data distribution, maintaining the structural integrity of the dataset, its robust nature, and its resilience to outlier influence, as cited in existing literature [57]. Through the application of mean imputation, the objective was to systematically and credibly address the gap of missing data, thereby enabling a more exhaustive examination of patient data and aiding in the construction of precise predictive models for the diagnosis of sepsis. The selection of this particular method was influenced by its capability to equilibrate data integrity with computational practicability, thereby enhancing the overall trustworthiness of the study's outcomes and derived conclusions.

3.6. Descriptive Statistics

The import of descriptive statistical techniques as essential tools for characterizing the inherent qualities of particular datasets has been emphasized by empirical research. These numerical methods enable an exhaustive examination of the fundamental features of data dispersion, thus providing a holistic grasp of its intrinsic attributes.

Descriptive statistics is a branch of statistics that provides summary details about data, which can be either a representation of the entire population or a sample of it. These details are mathematical summaries of the data and include the following measures:

Mean is the average value of a data set, calculated by summing all data points and dividing by the number of data points [57].

$$\mu = \frac{\sum_{i=1}^{N} X_i}{\mathrm{N}}$$

μ: Mean

N: Number of observations

X_i: i^{th} observation in the population

Standard Deviation measures the amount of variation or dispersion in a set of values. A low standard deviation indicates that the values tend to be close to the mean, while a high standard deviation indicates that the values are spread out over a wider range [58].

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

σ : Standart derivation

X_i: i^{th} observation in the population

μ: Mean

N: Number of observations

Variance is a statistical measurement of the spread between numbers in a data distribution. It squares the standard deviation, thus always providing a non-negative number.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

$\boldsymbol{\sigma}^2$: Variance

X_i: i^{th} observation in the population

μ : Mean

N: Number of observations

Table 15 illustrates the descriptive statistical analysis pertaining to the cohorts of infant age groups.

Table 15. The descriptive statistics of infant cohort

Infant Cohort					
Descriptive Statistics	Mean	Standart Deviation	Variance	Minimum	Maximum
age	0,00	0,02	0,00	0,00	0,39
heartrate	145,86	16,97	288,03	24,00	252,00
resprate	NA	NA	NA	NA	NA
temp	37,42	0,73	0,54	35,90	38,80
sysbp	NA	NA	NA	NA	NA
spo2	NA	NA	NA	NA	NA

diasbp	NA	NA	NA	NA	NA
meanbp	NA	NA	NA	NA	NA
wbc	14,61	6,47	41,83	1,00	60,90
bun	11,19	4,82	23,25	4,00	28,00
creatinine	0,60	0,31	0,10	0,20	1,50
ph	7,29	0,10	0,01	6,69	7,57

Table 15 presents the descriptive statistics of various health parameters for an infant cohort. Each row of the table represents a different health parameter, and the columns give the mean, standard deviation (SD), variance, and minimum and maximum values for each parameter.

Age: The mean age in this cohort is 0.00, indicating that this data represents newborn infants. The SD and variance are also close to zero, suggesting the ages are closely grouped around the mean. The maximum age, however, is 0.39 which, suggests some infants are around 4.5 months old.

Heart rate: The mean heart rate is 145.86 beats per minute, which is within the typical range for infants. The SD of 16.97 and variance of 288.03 suggests some variability in the heart rates within this group, but still within a reasonable range for infants.

Respiratory rate, *systolic blood pressure*, *oxygen saturation* (spo2), *diastolic blood pressure* (diasbp), and *mean arterial pressure* (meanbp) are all recorded as NA, indicating that this data is not available for this cohort.

Body Temperature: The average body temperature is 37.42 degrees Celsius, with a small SD and variance, suggesting most infants in this group have a body temperature close to the average.

White blood cell count (wbc): The mean is 14.61 $x10^3/\mu L$, which falls within the wide reference range for infants. The SD and variance suggest a reasonable variability in WBC counts in this cohort.

Blood Urea Nitrogen (bun): The mean BUN is 11.19 mg/dL, which is within the normal range for infants. The SD and variance values suggest some variation among the individuals in this cohort.

Creatinine: The mean level is 0.60 mg/dL, which falls within the normal range for infants. The SD and variance values suggest a reasonable level of variation among the individuals in this cohort.

Blood pH: The mean blood pH is 7.29, slightly lower than the average adult range, which can be common in newborns. The SD and variance values suggest a small degree of variation among individuals in this cohort.

Table 16 illustrates the descriptive statistical analysis pertaining to the cohorts of elder age groups.

Table contains descriptive statistics for a variety of health parameters in an elder cohort. Each row of the table corresponds to a different health parameter. The columns give the mean, SD, variance, minimum and maximum for each parameter.

Table 16. The descriptive statistics of elder cohort

Elder Cohort					
Descriptive Statistics	Mean	Standard Deviation	Variance	Minimum	Maximum
age	75,21	9,08	82,53	60,00	91,4
heartrate	85,88	19,61	384,73	23,00	173
resprate	19,33	5,55	30,79	4,00	57
temp	36,59	0,97	0,95	26,11	41,1
sysbp	123,09	24,57	603,71	34,00	263
spo2	96,59	5,07	25,72	1,00	100
diasbp	60,66	15,00	225,01	10,00	177
meanbp	79,50	16,47	271,15	14,00	188
wbc	12,82	10,53	110,85	0,10	326
bun	31,02	23,43	548,94	3,00	252
creatinine	1,53	1,41	1,98	0,10	15,9
ph	7,36	0,10	0,01	6,64	7,72

Table 17 illustrates the descriptive statistical analysis pertaining to the cohorts of all age groups.

Table 17 . The descriptive statistics of all age cohort

All Age Cohort					
Descriptive Statistics	Mean	Standart Deviation	Variance	Minimum	Maximum
age	55,64	27,01	729,26	0,00	91,40

heartrate	98,64	29,59	875,74	23,00	252,00
resprate	19,14	5,65	31,92	4,00	61,00
temp	36,67	0,99	0,97	26,11	42,00
sysbp	123,14	23,92	571,96	34,00	263,00
spo2	96,82	4,89	23,93	1,00	100,00
diasbp	63,27	15,66	245,09	10,00	182,00
meanbp	81,27	16,83	283,33	2,00	194,53
wbc	13,30	9,18	84,27	0,10	326,00
bun	28,06	23,29	542,29	1,00	252,00
creatinine	1,53	1,68	2,83	0,10	28,60
ph	7,35	0,10	0,01	6,64	7,72

Table 17 provides various descriptive statistics for different health indicators of all age cohort.

Age: The mean age is 55.64 with a standard deviation of 27.01 and a variance of 729.26. This high standard deviation and variance imply a wide range of ages in the data set.

Heart Rate: The mean heart rate is 98.64 with a standard deviation of 29.59 and a variance of 875.74. This high standard deviation and variance indicate a wide spread in heart rates among the population.

Respiratory Rate: The mean is 19.14, the standard deviation is 5.65, and the variance is 31.92. The values show some degree of variability in the respiratory rates but not as high as the heart rates.

Body Temperature: The mean body temperature is 36.67°C, with a standard deviation of 0.99 and variance of 0.97. The small standard deviation and variance indicate that most individuals have a body temperature close to the mean.

Systolic Blood Pressure: The mean is 123.14, the standard deviation is 23.92, and the variance is 571.96. These numbers suggest that the systolic blood pressure among the individuals varies to a significant degree.

Oxygen Saturation: The mean is 96.82%, with a standard deviation of 4.89 and variance of 23.93. These indicate a moderate variability in oxygen saturation levels.

Diastolic Blood Pressure: The mean is 63.27, the standard deviation is 15.66, and the variance is 245.09. Similar to systolic blood pressure, the diastolic blood pressure varies notably among the cohort.

Mean Arterial Pressure: The mean is 81.27, the standard deviation is 16.83, and the variance is 283.33. These numbers again show a considerable spread in the values.

White Blood Cell Count: The mean is 13.30, the standard deviation is 9.18, and the variance is 84.27. This implies a high degree of variability in the WBC counts among the population.

Blood Urea Nitrogen: The mean is 28.06, the standard deviation is 23.29, and the variance is 542.29. These show a high variability in BUN levels.

Creatinine: The mean is 1.53, the standard deviation is 1.68, and the variance is 2.83. These values indicate a considerable variability in creatinine levels.

Blood pH: The mean is 7.35, the standard deviation is 0.10, and the variance is 0.01. These values suggest a low variability in blood pH levels.

3.7. Chi-Square Test

Employing the Chi-square test, an exploration was conducted to ascertain the existence of a potential correlation between attribute and label variables across all cohorts. The value, denoting the benchmark for statistical significance, was established at 0.01, conforming to the conventional criteria typically utilized in diagnostic procedures associated with disease [59]. The consequent P values, derived from this evaluation, are encapsulated in Table 18, correspondingly designated for the relevant admission time period.

Table 18. Chi-Square test results

Attribute	**All Age Cohort**	**Elder Cohort**	**Infant Cohort**
Age	8,46E-77	0.000127	0.000127
Heart rate	2,39E-57	>0.01	>0.01
Respiratory rate	1,92E-25	>0.01	>0.01
Temperature	1,02E-10	>0.01	>0.01
Systolic Blood Pressure	1,32E-08	>0.01	>0.01
Oxygen Saturation	6,30E-03	>0.01	>0.01
Diastolic Blood Pressure	6,29E-15	>0.01	>0.01
Mean Arterial Blood Pressure	3,58E-14	>0.01	>0.01
White Blood Cell Count	>0.01	>0.01	>0.01
Blood Urea Nitrogen	7,92E-42	>0.01	>0.01
Creatinine	5,38E-63	5,11E-10	5,11E-10
Arterial pH	>0.01	>0.01	>0.01

Date and time of patients' entry in the ICU	>0.01	>0.01	>0.01
Date and time of patients' discharge from ICU	>0.01	>0.01	>0.01
Time of Suspected Infection	0.00191	>0.01	>0.01

Table 18 presents p-values from a Chi-square test for several health attributes across three different cohorts: All Age, Elder, and Infant. Given that the threshold for significance (alpha level) is 0.01, any p-value lower than 0.01 means that rejecting the null hypothesis and considering the result statistically significant. That is, it will be assumed there's an association between the cohort and the health attribute. Conversely, a p-value higher than 0.01 means that null hypothesis is not refejected and conclude that there's no evidence of an association.

Age: The p-values are below 0.01 for all cohorts, meaning there's a statistically significant association between age and the three cohorts. It's not surprising since age is the defining variable of these cohorts.

Heart rate: There's a significant association for the All Age Cohort, but not for the Elder and Infant cohorts. This means that while heart rate varies significantly across all ages, it doesn't show significant variation within the Elder and Infant cohorts.

Respiratory rate, *temperature*, *systolic blood pressure*, *oxygen saturation*, *diastolic blood pressure*, *mean arterial blood pressure*: These all show the same pattern, a significant association for the All Age Cohort but not for the Elder and Infant cohorts. This could indicate that while these parameters vary across all ages, they are relatively constant within the Elder and Infant Cohorts.

White blood cell count, *blood pH*, and the *date and time of patients' entry in the ICU and discharge from ICU*: The p-values are above 0.01 for all cohorts, indicating no significant association with the cohort.

Blood Urea Nitrogen and *Creatinine*: For these attributes, there's a significant association for the All Age Cohort, and also for the creatinine attribute within the Elder and Infant Cohorts.

Time of Suspected Infection: This attribute has a significant association with the All Age Cohort, but not with the Elder and Infant Cohorts.

3.8. Diagnostic Test

Within the purview of this research, machine learning models were established for performance appraisal predicated upon the sepsis status of patients, gauged via SOFA scores. The International Classification of Diseases, Ninth Revision (ICD-9) codes 99592, 99591, and 78552 were leveraged to categorize patients from the MIMIC III dataset, the primary data corpus of this study, as sepsis-positive instances [60]. The utility of ICD codes lies in their facilitation of clinical evaluation and diagnostic procedures as they streamline the organization of medical records, enable the estimation of disease prevalence, and aid in statistical analysis. A comprehensive scrutiny of the correlation between the ICD code and the SOFA score during the model construction phase provides valuable insights into the model's practical applicability and validation, fortifying clinical decision-making processes. Table 19 delineates the evaluation criteria for the SOFA score, serving as an indicator of sepsis presence.

Table 19. Classification model evaluation

Classification Model Evaluation Criteria (ICD9vsSOFA)	**All Age**	**Elder**	**Infant**
Sensitivity	0,932	0,940	0,667
Specificity	0,487	0,403	0,723

False Positive Rate	0,513	0,597	0,277
False Negative Rate	0,068	0,060	0,333
Prevalence	0,138	0,173	0,001
Positive Predictive Value	0,226	0,248	0,002
Negative Predictive Value	0,978	0,970	1,000

Table 19 presents different evaluation metrics for a classification model comparing ICD9 versus SOFA scores across three cohorts: All Age, Elder, and Infant.

	ICD 9 Code Sepsis Present	ICD 9 Code Sepsis Absent	
SOFA Score Sepsis Positive	True Positives (TP)	False Positives (FP)	TOTAL SOFA SCORE SEPSIS POSITIVE (TP + FP)
SOFA Score Sepsis Negative	False Negatives (FN)	True Negatives (TN)	TOTAL SOFA SCORE SEPSIS NEGATIVE (FN + TN)
	TOTAL SEPSIS (TP + FN)	TOTAL NORMAL (FP + TN)	TOTAL POPULATION (TP + FN + FP + TN)

Sensitivity: Sensitivity is the ability of a test to correctly classify an individual as "*diseased*" [61].

$$Sensitivity = \frac{TP}{TP + FP}$$

Also known as the True Positive Rate, sensitivity measures the proportion of actual positives (cases with condition) that are correctly identified as such. Higher sensitivity means the model is good at catching positives. For All Age, Elder, and Infant Cohorts, the sensitivity values are 0.932, 0.940, and 0.667 respectively. This suggests the model

performs well at identifying true positives for the All Age and Elder Cohorts, but less so for the Infant Cohort.

Specificity: The ability of a test to correctly classify an individual as *disease- free* is called the test′s specificity [62].

$$Specificity = \frac{TN}{TN + FP}$$

Specificity is the True Negative Rate, i.e., the proportion of actual negatives (cases without the condition) that are correctly identified. Higher specificity means the model is good at confirming negatives. The model has low specificity for All Age (0.487) and Elder (0.403) cohorts, and moderate for Infant (0.723). This implies it's not as effective at identifying true negatives in the All Age and Elder cohorts.

False Positive Rate (FPR): This is the proportion of actual negatives that are incorrectly identified as positives. Lower FPR is desirable. The model's FPR is quite high for the All Age (0.513) and Elder (0.597) cohorts and moderate for the Infant (0.277) cohort. It's the probability that a false alarm will be raised: that a positive result will be given when the true value is negative.

$$FPR = \frac{FP}{FP + TN}$$

False Negative Rate (FNR): The false negative rate – also called the miss rate – is the probability that a true positive will be missed by the test. This is the proportion of actual positives that are incorrectly identified as negatives. Lower FNR is desirable. The model shows a low FNR for the All Age (0.068) and Elder (0.060) cohorts and higher for the Infant cohort (0.333).

$$FNR = \frac{FN}{FN + TP}$$

Prevalence: This refers to the actual occurrence of the condition in the population. In this table, prevalence is significantly higher in the Elder cohort (0.173) compared to the All Age (0.138) and Infant (0.001) cohorts.

$$Prevalance = \frac{TP + FN}{FN + TP + FP + TN}$$

Positive Predictive Value (PPV): It is the percentage of patients with a positive test who actually have the disease [63].

$$PPV = \frac{TP}{TP + FP}$$

Also known as Precision, PPV is the proportion of positive results that are true positives. Higher PPV is desirable. The model has low PPV for all cohorts, with the highest being the Elder cohort at 0.248.

Negative Predictive Value (NPV): It is the percentage of patients with a negative test who do not have the disease.

$$NPV = \frac{TN}{FN + TN}$$

This is the proportion of negative results that are true negatives. Higher NPV is desirable. The model exhibits high NPV for all cohorts, particularly for the Infant cohort, where it reaches 1.000, implying a perfect score.

In summary, it performs best in the Elder Cohort with higher sensitivity and NPV. However, its specificity and PPV are generally low across all cohorts. It seems to struggle particularly with the Infant Cohort, where it misses a significant proportion of actual positives (as evidenced by the sensitivity and FNR values).

An elevated sensitivity value holds particular merits for conditions such as sepsis, which demand immediate medical attention. Sensitivity, equivalently denoted as the true positive rate, plays a pivotal role in the precise identification of individuals afflicted with sepsis. Consequently, a considerable segment of sepsis patients are expected to be appropriately classified owing to enhanced sensitivity, a key element instigating the commencement of expeditious intervention and therapeutic oversight.

Observed sensitivity values stood at an elevated 0.932 for the all age group, 0.940 for the subgroup of elderly individuals, and 0.667 for the infant subgroup, evident across each of the cohorts. These markedly sensitive measures suggest that a substantial fraction of sepsis patients within each cohort were aptly classified. Further, to enhance the robustness and efficacy of the evaluation procedure, the labels assigned to the patients contingent on their SOFA scores were factored into the appraisal of the machine learning models.

The research evaluated the efficacy of CNN, LSTM, MLP, RF, and XGBoost algorithms applied to hourly datasets from three patient cohorts. Datasets were partitioned into an 80-20 split, assigning the larger segment for training and the smaller for testing purposes. The model's performance was appraised utilizing the 5-fold cross-validation technique, invoking performance metrics such as accuracy, precision, AUC, F1 score, sensitivity, and specificity.

Accuracy: This is the ratio of the correctly predicted instances to the total instances. It measures the overall correctness of the model.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision: Also known as the Positive Predictive Value, it is the ratio of correctly predicted positive instances to the total predicted positive instances. It measures the model's ability to correctly identify only relevant instances [64].

$$Precision = PPV = \frac{TP}{TP + FP}$$

Area Under the Curve (AUC): The AUC refers to the area under the ROC curve, which plots the TPR (Sensitivity) against the FPR (1-Specificity) at various threshold settings [65]. The AUC provides an aggregate measure of performance across all possible classification thresholds. An AUC of 1 represents a perfect model.

F1 Score: The F1 score is the harmonic mean of precision and recall (sensitivity), where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0 [66].

$$F1 = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall}$$

Sensitivity (Recall or TPR): This is the ratio of correctly predicted positive instances to the actual positive instances. It measures the model's ability to correctly identify all relevant instances.

$$Recall = Sensitivity = TPR = \frac{TP}{TP + FN}$$

Specificity (TNR): This is the ratio of correctly predicted negative instances to the actual negative instances. It measures the model's ability to correctly identify all non-relevant instances.

$$Specificity = TNR = \frac{TN}{TN + FP}$$

4. IMPLEMENTED METHODS

4.1. Multi-layer Perceptron (MLP)

In contemporary computational science, a neural network is conceptualized as a sophisticated computational paradigm meticulously engineered to discern intricate patterns within extensive datasets. This model, through rigorous training on ample instances, refines its capacity to execute designated tasks with enhanced efficiency compared to traditional algorithmic approaches. Within the broad spectrum of neural networks, the MLP emerges as a distinctive subset. It is characterized by an assembly of interlinked perceptron, which can be formalized as mathematical constructs. Each inter-perceptron connection possesses a quantifiable strength, termed "*weight*", which plays a pivotal role in determining the cumulative output of the network. It is worth noting that MLPs are equivalently recognized in the literature as feedforward artificial neural networks [67].

4.1.1. Neural Networks

Neural networks serve as computational paradigms, deriving their foundational principles from the intricate architecture of the cerebral system, albeit not endeavoring to replicate it with exactitude. Historically, the human brain, a marvel of complex cognitive processing, has ignited curiosity and has been an anchor for exploration across diverse scientific realms. While certain incarnations of neural networks aspire to elucidate cerebral functionalities, it is imperative to note that state-of-the-art DL methodologies refrain from mirroring the precise cerebral computations. Instead, the primary objective of DL is the formulation of systems proficient in discerning and assimilating multifarious patterns [68].

In a landmark collaboration during the 1940s, neurophysiologist Warren McCulloch and logician Walter Pitts pioneered a seminal model attempting to emulate specific facets of cerebral activity. Their devised model, designated as a "*neuron*", operated linearly, deducing either affirmative or negative outcomes contingent on designated inputs and associated weights.

$$f(x, w) = x_1 w_1 + \cdots + x_n w_n$$

This computational paradigm was conceptualized to emulate the intrinsic operations of the brain's quintessential unit, the neuron. Drawing parallels to the manner in which cerebral neurons propagate electrical impulses, the model devised by McCulloch and Pitts processed inputs and conveyed them to interconnected neurons, contingent upon the magnitude and intensity of these signals.

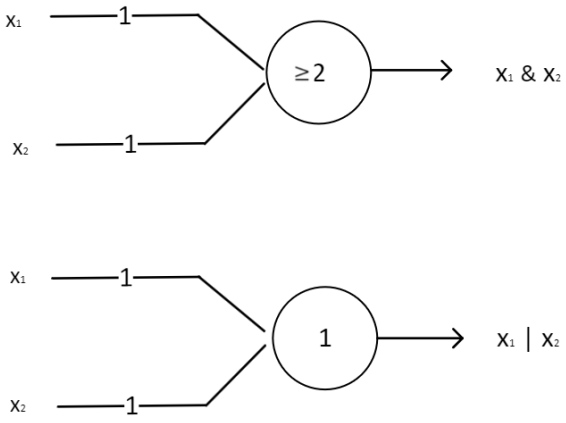

Figure 2 Neurons

The preliminary model proposed by McCulloch and Pitts was capable of emulating a logical gate, processing singular or dual binary inputs to yield an output governed by a Boolean function. This output was contingent upon specific input-weight configurations. However, a notable constraint of this model was its static learning capacity; the outcome was determinable only if the weights, serving as modulators, were pre-specified.

The proposition that "*the nervous system is comprised of neurons, each characterized by a soma and an axon, with a requisite excitation threshold for neuronal impulse transmission*" elucidates the complexities inherent in neural operations [69].

Subsequent to the foundational contributions of McCulloch and Pitts, Frank Rosenblatt furthered this domain by introducing the Perceptron in the subsequent decade. This innovation facilitated the algorithm's capacity for dynamic learning, empowering it to autonomously modify the weights to ascertain the anticipated output.

4.1.2. Perceptron

The Perceptron, initially envisioned as an image recognition apparatus, is contemporarily recognized predominantly as an algorithm. The nomenclature "*perceptron*" is derived from its capacity to emulate human perceptual abilities, predominantly in visual image discernment.

The underlying ambition for such an apparatus was its capability to directly assimilate inputs from the ambient environment, which includes sensory stimuli such as light, acoustic waves, and thermal variations — essentially capturing the entirety of the perceivable realm or the "*phenomenal world*". This direct assimilation negated the necessity for human-mediated data processing and encoding [70].

Frank Rosenblatt's perceptron apparatus was predicated upon a quintessential computational entity, referred to as the neuron. In alignment with antecedent models, each neuron in Rosenblatt's architecture encompassed a cellular structure primed to receive an amalgamation of inputs juxtaposed with their respective weights.

A salient feature of Rosenblatt's neuron was the modality of input processing. Subjected inputs experienced a weighted summation. Subsequent to this computation, if the cumulative value transcended a pre-established threshold, the neuron would exhibit activation or "*firing*", culminating in an output production.

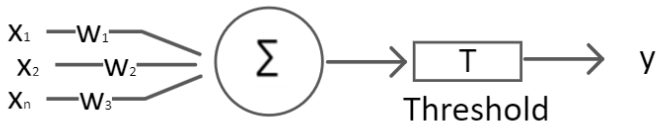

Figure 3. Activation Function

$$y = \begin{cases} 1, & if \sum_i w_i x_i - T > 0 \\ 0, & otherwise \end{cases}$$

This demarcated threshold, represented as T, functioned intrinsically as the activation function. When the aggregate of weighted inputs exceeded the value of zero, the neuron would elicit an output of 1. Conversely, if this condition was unmet, the output rendered would be zero.

4.1.2.1. Perceptron for Binary Classification

The perceptron is designed to produce a discrete output, governed by its activation function, thus facilitating its role as a binary classification model. This particular model delineates a linear decision boundary, with the intent of identifying a hyperplane that minimizes the discrepancy between the decision boundary and erroneously classified data points [71].

$$D(w, c) = -\sum_{i \in M} y_i(x_i w_i + c)$$

For optimization purposes, the Perceptron employs the Stochastic Gradient Descent methodology. In scenarios where data are linearly separable, the convergence of Stochastic Gradient Descent is guaranteed within a finite number of iterations.

The activation function plays a pivotal role in determining the neuron's activation state. Historically, perceptron models incorporated the sigmoid function as their chosen

activation function. This function possesses the inherent ability to map any given real number input to an output range between 0 and 1, embodying a non-linear relationship. This ensures that even with potential negative values as input, the resulting neuron output is binary, either 0 or 1.

$$f(x) = \frac{1}{1 + e^{-x}}$$

However, in recent years, there has been a discernible shift in Deep Learning methodologies, exhibiting a predilection for the Rectified Linear Unit (ReLU) as the preferred activation function.

$$f(x) = \max(0, x)$$

The rising prominence of ReLU can be attributed to its synergistic relationship with Stochastic Gradient Descent, computational expediency, and its property of scale-invariance. During the input processing phase in the neuron, an initial set of weights is selected at random. Subsequent to a weighted summation of these weights, the activation function, typically the ReLU in contemporary applications, determines the resultant output value.

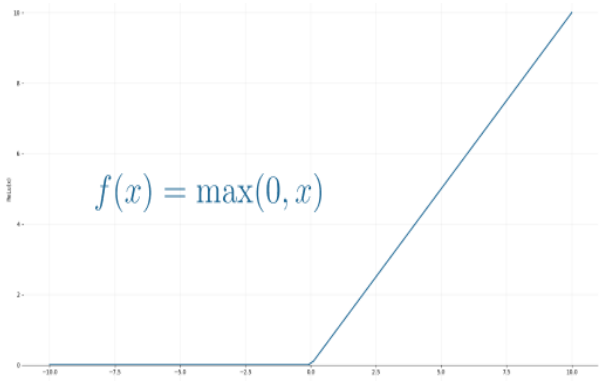

Figure 4. ReLU Function

The Perceptron employs Stochastic Gradient Descent to identify the optimal weight set that reduces the distance between the decision boundary and the misclassified data points.

Upon achieving convergence through Stochastic Gradient Descent, the data is partitioned by a linear hyperplane.

Contrary to initial proclamations that the Perceptron could simulate any circuit or logical function, it encountered critique for its incapacity to represent the XOR (exclusive OR) gate. The XOR gate produces an output of 1 exclusively when there is a disparity in inputs. This constraint was corroborated by Minsky and Papert in 1969 [72], emphasizing that a Perceptron with a singular neuron is ineffective when addressing non-linear data.

4.1.3. Function in feed forward neural network

4.1.3.1. Cost function

In feedforward neural networks, the cost function holds paramount significance as it quantifies the divergence between predicted and true values. Marginal modifications in the weights and biases might exert insubstantial influence on the designated data points. Therefore, a continuous cost function is integrated to ascertain the preeminent methodology for adjusting weights and biases, thereby enhancing the accuracy of the network. A prevalent cost function employed for this purpose is the mean square error (MSE), articulated as:

$$C(w,b) \equiv \frac{1}{2n}\sum_{x} \| \, \mathrm{y(x)} - \mathrm{a} \, \|^2$$

Where:

W: Delineates the weights within the network.

b: Symbolizes biases.

n: Corresponds to the aggregate count of training inputs.

a: Represents the output vector.

x: Is indicative of the input.

‖v‖: Signifies the norm or magnitude of vector v.

It is pertinent to note that, in the presented information, the symbol for *biases* (b) was elucidated but remained unincorporated within the cost function. Concurrently, although the function made a reference to the norm of the vector, it was not explicitly included in the preliminary equation; nonetheless, it has been amalgamated in the aforementioned equation [73].

4.1.3.2. Loss function

Within the realm of neural networks, the loss function is of paramount importance in assessing the imperative for adjustments throughout the learning process. The quantity of neurons in the output layer is commensurate with the number of distinct classes, highlighting the discrepancies between the anticipated and genuine probability distributions [74]. In cases of binary classification, the cross-entropy loss can be articulated as:

Cross Entropy Loss:

$$L(\Theta)\begin{cases}-\log(\hat{y}) & if\ y=1\\ -\log(1-\hat{y}) & if\ y=0\end{cases}$$

4.1.4. Gradient learning algorithm

In the context of the gradient descent algorithm, the next iteration point is ascertained by amending the gradient of the current position, modulated by a designated learning rate. This revision necessitates the subtraction of the result from multiplying the gradient with the learning rate from the extant position. To minimize the function's objective value, the product is deducted from the current point. The update mechanism can be formally represented as:

$$p_{n+1} = p_n - \eta\nabla f(p_n)$$

In this expression, η signifies the learning rate, which governs the step magnitude within the algorithmic process. The appropriate selection of the learning rate is of critical importance in machine learning, given its substantial impact on the model's efficacy.

4.1.5. Output units

Within a neural network architecture, the output layer comprises output units tasked with producing the expected outcome or prediction, thereby facilitating the realization of the network's designated objective. The choice of these output units is intrinsically associated with the decision pertaining to the cost function. It is noteworthy that any unit which operates as a hidden unit within the network's framework can also serve in the capacity of an output unit.

The Multilayer Perceptron (MLP), a sophisticated augmentation to the feedforward artificial neural network paradigm, was formulated to surmount the inherent limitations of linear neural architectures. Distinctively, the MLP is engineered to yield a series of outputs predicated upon designated inputs. This network architecture is delineated by several layers of nodes, systematically structured in a directed acyclic graph, thereby guaranteeing an exclusively unidirectional propagation of data from the input nodes towards the terminal output nodes. The capacity for non-linear mapping between input and output nodes renders the MLP invaluable for computational tasks that defy linear separability. Notably, the primary application domain of MLPs remains within the realm of supervised learning paradigms. Their efficacy is manifest in a myriad of applications, including but not limited to, speech and image recognition, machine translation, and they additionally serve as pivotal instruments in advanced research pursuits in computational neuroscience and parallel distributed processing [75]. At its core, the MLP's architecture, complemented by its inherent non-linearity, equips it to approximate a broad spectrum of continuous functions.

In its structural taxonomy, the MLP comprises three principal tiers: the initial input layer, tasked explicitly with assimilating external signals; an intermediary layer or layers, colloquially termed hidden layers, entrusted with intricate computational responsibilities;

and a culminating output layer, meticulously calibrated for intricate operations such as data categorization and predictive analytics. It is imperative to underscore that, with the exception of input nodes, all nodes in the MLP architecture are intrinsically bound to a non-linear activation function. As a corollary, whereas foundational Perceptron models necessitate neurons to harness threshold-centric activation paradigms, such as the Rectified Linear Unit (ReLU) or sigmoid function, the MLP framework provides the latitude for neurons to assimilate any conceivable non-linear activation function.

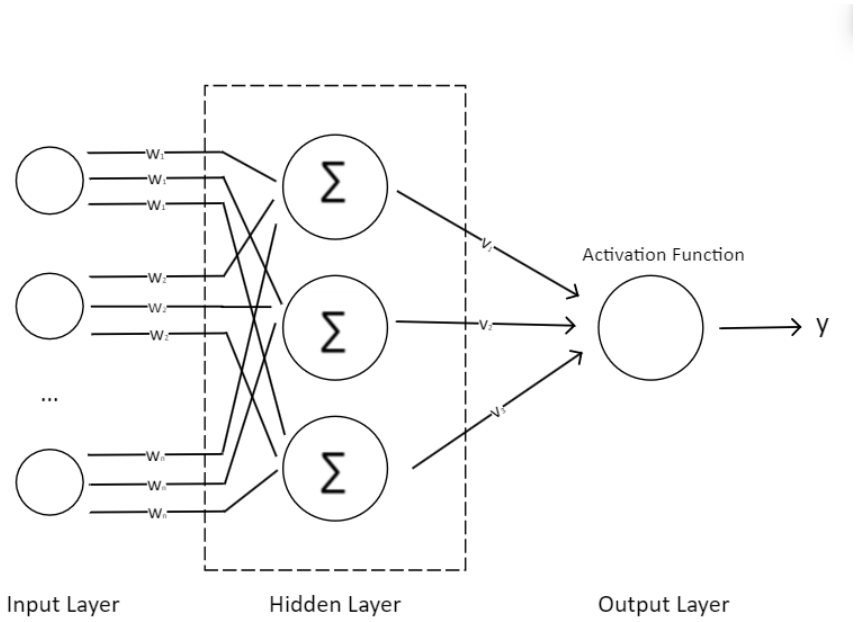

Figure 5. MLP Structure

Backpropagation stands as the preeminent supervised learning algorithm for the optimization of MLPs. Given its intricately layered neuronal architecture, the MLP can be aptly categorized under the umbrella of deep learning methodologies. In encapsulating its essence, as elucidated by Tamouridou et al., the MLP emerges as a formidable feed-forward neural network mechanism, characterized by its unidirectional data flow, commencing from the input stratum and culminating seamlessly at the output tier.

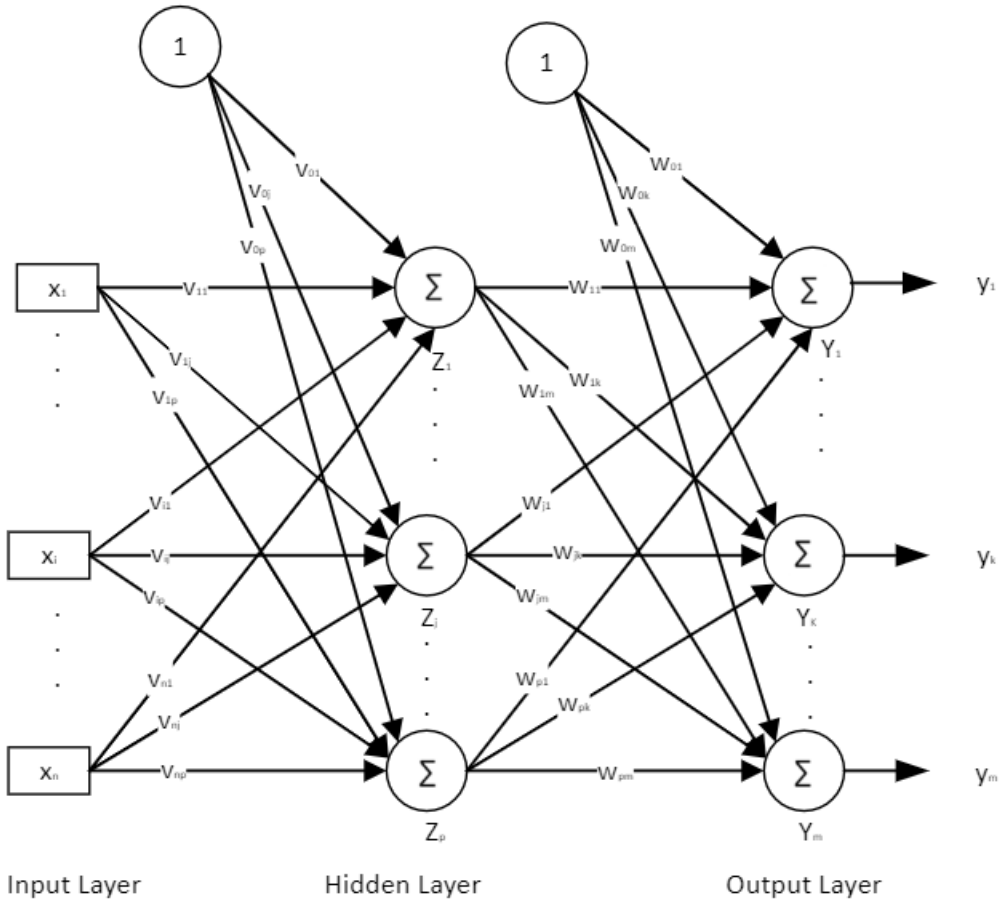

Figure 6. MLP Architecture

Within the architecture of the MLP, each node undergoes training via the backpropagation learning algorithm, underscoring the network's adeptness in addressing intricate computational challenges. The computations executed by every neuron, predominantly within the hidden and output layers, constitute the foundational operations of the MLP.

The MLP is categorized as a feedforward algorithmic structure. Analogous to the foundational Perceptron model, the MLP integrates inputs using initial weightings in a linear combination, subsequently channeling this amalgamation through an activation function. A salient characteristic that delineates the MLP from the single-layer Perceptron is its systematic progression, directing each linear output to successive neuronal layers.

Within this architectural paradigm, each respective layer computes an intrinsic data representation, relaying its computational outcomes to the succeeding layer. This ordered feedforward progression persists through the myriad hidden layers until it culminates at the output stratum.

Nevertheless, the learning paradigm is not merely confined to the computation of weighted linear combinations and their propagation. Absent iterative refinements, a simplistic approach of singular computation and relay to the output layer would be insufficient in optimizing the weights to minimize a designated cost function. Thus, the learning mechanism must be recognized as an iterative endeavor; a sole iteration would be inadequate for the algorithm to effectively refine its weights based on the presented dataset.

4.1.6. Backpropagation

The backpropagation algorithm acts as the cardinal technique for iteratively refining the weight parameters within the Multilayer Perceptron. The fundamental aim of this iterative weight adjustment is the minimization of a specified cost function. For the effective operation of backpropagation, certain conditions must be satisfied: the functions engaged in the synthesis of inputs and associated weights, exemplified by the linear weighted sum, along with the activation functions such as the ReLU, must exhibit differentiable characteristics. Additionally, the derivatives of these functions ought to remain within bounded limits.

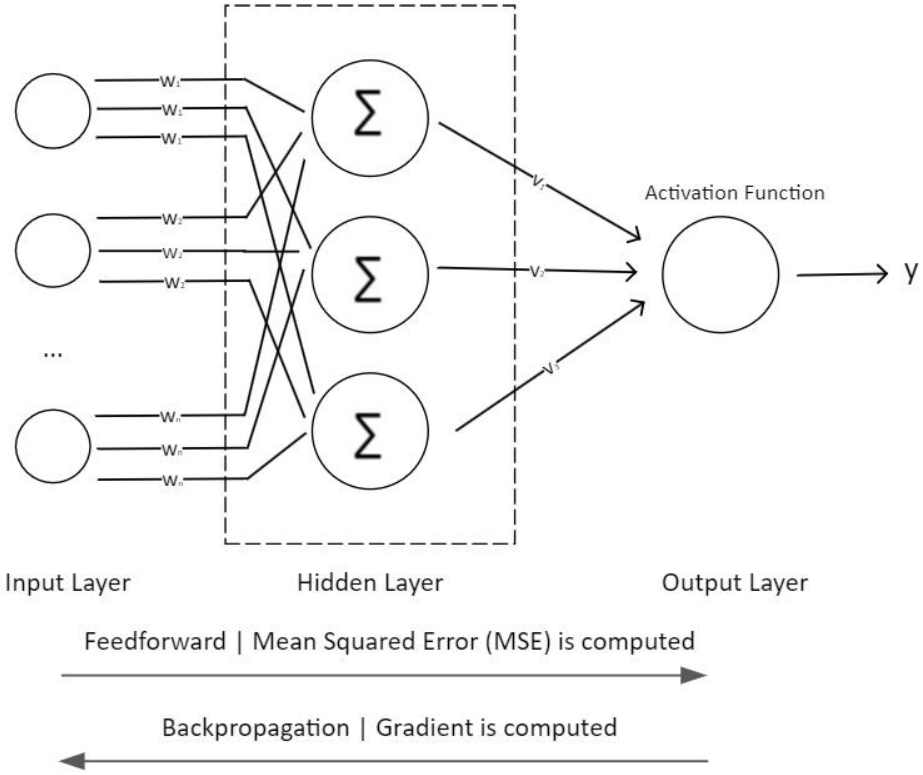

Figure 7. Backpropagation and Feedforward

The imperative for derivatives to remain within specified bounds stems from the prevalent use of the Gradient Descent optimization technique in the MLP. In each iterative cycle, subsequent to the computation and forward propagation of weighted sums across all layers, the gradient of the MSE with respect to each weight parameter is ascertained for the entirety of the input-output pairs. This derived gradient then informs the subsequent modification of the weights associated with the inaugural hidden layer. This systematic process of weight recalibration, commencing from the output layer and retrogressing through to the input layer, is emblematic of the "*backpropagation*" methodology.

$$\Delta w(t) = -\epsilon \frac{dE}{dw(t)} + \alpha \Delta w(t-1)$$

The optimization process persists in a cyclical manner until the gradient for each input-output pair reaches a state of convergence, which is indicated when the alteration in the gradient across consecutive iterations remains inferior to a predetermined convergence criterion.

The MLP stands as a seminal construct within the landscape of neural network architectures. An exhaustive mathematical exploration of its constituent elements is pivotal to grasp its operational intricacies in their entirety [70].

4.1.7. Pseudocode

```
Procedure MLP_Training(data, labels, epochs, learning_rate):
    Initialize weights W and biases b for all layers randomly
    For epoch in 1 to epochs:
        For each (input, target) in (data, labels):
            forward_pass(input)
            compute_loss(target)
            backpropagate_loss()
            update_weights_and_biases(learning_rate)
        EndFor
    EndFor

Procedure forward_pass(input):
    a[0] = input
    For l in 1 to L: // L is the number of layers
        z[l] = W[l] * a[l-1] + b[l]
        a[l] = activation_function(z[l])
```

```
    EndFor
    return a[L]

Procedure compute_loss(target):
    loss = loss_function(a[L], target)
    return loss

Procedure backpropagate_loss():
    delta[L] = loss_derivative(a[L], target) * activation_derivative(z[L])
    For l in L-1 down to 1:
        delta[l] = (W[l+1].transpose() * delta[l+1]) * activation_derivative(z[l])
    EndFor
    For l in 1 to L:
        gradient_W[l] = delta[l] * a[l-1].transpose()
        gradient_b[l] = delta[l]
    EndFor

Procedure update_weights_and_biases(learning_rate):
    For l in 1 to L:
        W[l] = W[l] - learning_rate * gradient_W[l]
        b[l] = b[l] - learning_rate * gradient_b[l]
    EndFor
```

4.2. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) represents a specialized variation of the RNN architecture, explicitly engineered to counteract the inherent difficulties posed by the sequential data processing, encompassing domains such as temporal series, auditory speech signals, and linguistic texts. While conventional RNNs operate by recycling the output from an antecedent step as an input for the subsequent one, they often falter in preserving extended temporal dependencies inherent within data sets. This limitation invariably compromises their operational efficiency, particularly when confronted with expansive temporal intervals interlacing pertinent information. Conversely, LSTMs manifest an intrinsic aptitude to sustain information across protracted timeframes, which invariably augments their proficiency in forecasting and categorization tasks predicated upon temporal series data.

LSTM, as proposed by Hochreiter & Schmidhuber, was conceived as an advancement to surmount the inherent constraints exhibited by traditional RNNs. The distinguishing feature of LSTMs resides in the incorporation of a unique memory cell, designed to retain information over extended temporal spans. The functionality of this memory cell is meticulously modulated by a triumvirate of gates: *the input, forget, and output gates.* Specifically, the input gate orchestrates the assimilation of information into the memory cell, the forget gate orchestrates the selective purging of information, and the output gate governs the dissemination of information to subsequent computational steps. This intricate configuration bestows upon LSTMs the capacity for discerning retention, elimination, and transmission of information, thereby enhancing their adeptness in discerning and leveraging extended temporal dependencies within sequential datasets [76].

Additionally, LSTM architectures can be hierarchically layered, culminating in the formulation of deep LSTM networks. This layered structuring augments their capacity to decipher and internalize increasingly complex patterns inherent in sequential datasets. Consequently, their sophisticated learning proficiencies render them especially apt for endeavors such as linguistic translation, auditory signal recognition, and temporal series prognostication.

4.2.1. Structure of LSTM

Within the framework of RNN, the LSTM emerges as a distinct architecture purposefully constructed to counteract the vanishing gradient dilemma frequently associated with traditional RNNs. Contrasting conventional RNNs, LSTM networks possess a singular internal configuration that adeptly facilitates the retention and manipulation of information across extended sequences [77, 78].

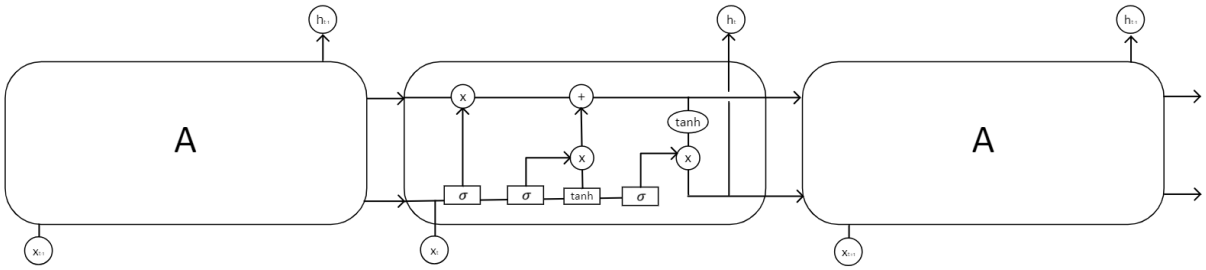

Figure 8. General Structure of LSTM

The structural design of the LSTM can be conceptualized as a sequential arrangement of memory units, termed cells, interlinked with four integral neural networks. The quintessential operations of the LSTM are coordinated via three principal components, commonly denoted as gates. These gates regulate the dissemination of data both internally within and externally between the memory units.

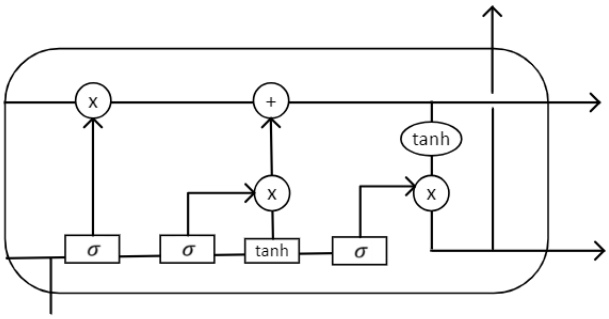

Figure 9. Structure of LSTM

4.2.1.1. Forget Gate

This gate is tasked with ascertaining the segments of information within the cell state that warrant retention or exclusion. For every data unit within the cell state, the forget gate yields an output in the continuum of 0 (indicative of total exclusion) to 1 (indicative of total retention). This determination is predicated upon the contemporary input, represented as x_t, coupled with the outcome emanating from the antecedent cell, depicted as h_{t-1} . This ensemble undergoes a multiplication operation with a weight matrix, subsequently amalgamating with a bias term. The ensuing value undergoes transformation via a sigmoid activation function to derive the output for the forget gate, formalized as:

$$f_t = \sigma(x_t * U_f + H_{t-1} * W_f)$$

Wherein:

W_f: Delineates the weight matrix specific to the forget gate.

h_{t-1}, x_t: Signifies the concatenation of the extant input with the prior hidden state.

Sigma: Demarcates the sigmoid activation function, circumscribing the output to the interval [0,1].

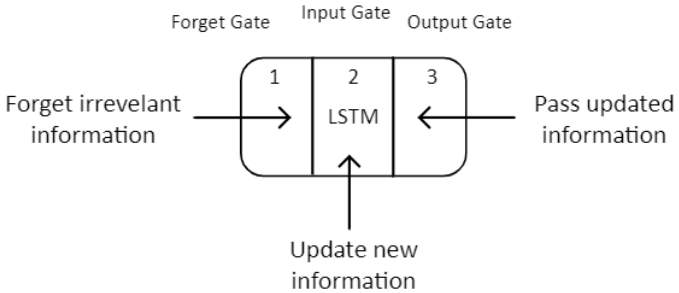

Figure 10. Gates of LSTM

Subsequent to this computation, f_t undergoes multiplication with the cell state from the preceding temporal point, thereby adjudicating the quantum of information to be preserved within the cell state for the succeeding computational phase.

$$C_{t-1} * f_t = 0 \quad \dots if\ f_t = 0\ (forget\ everthing)$$

$$C_{t-1} * f_t = 0 \quad \dots if\ f_t = 1\ (forget\ nothing)$$

4.2.1.2. Input Gate:

Input gate is entrusted with the task of appraising the relevance of incoming data vectors. This evaluation is executed by means of a specific mathematical formulation [79]:

$$i_t = \sigma(x_t * U_i + H_{t-1} * W_i)$$

Where:

X_t: Input at the current timestamp t.

U_i : Weight matrix of the input.

H_{t-1}: A Hidden state at the prior timestamp.

W_i : Weight matrix linked with the prior hidden state.

The sigmoid function ensures that the outcome i_t falls within a [0, 1] range.

$$N_t = \tanh(x_t * U_c + H_{t-1} * W_c)\ (new\ information)$$

Within the framework of Long Short-Term Memory networks, the synthesis of novel information destined for the cell state is derived from an antecedent hidden state, represented as h_{t-1}, and the contemporaneous input, denoted by x_t . The incorporation of the hyperbolic tangent (tanh) activation function ascertains that the resultant output lies within the bounds of -1 and 1. A value trending toward the positive spectrum intimates an augmentation to the cell state, whereas a negative inclination signifies a decrement.

It is pivotal to note that the aforestated information does not amalgamate directly with the cell state. It is subject to a subsequent procedural refinement:

$$C_t = f_t * C_{t-1} + i_t * N_t \ (updating\ cell\ state)$$

Where:

- C_{t-1}: Previous cell state.

- f_t, i_t,N_t: Previously computed values.

In the architecture of Long Short-Term Memory networks, the Input gate plays a pivotal role in modulating the incorporation of salient information into the cell state. Initially, the sigmoid activation function serves to regulate the information. During this phase, inputs h_{t-1} and x_t are subjected to a filtering mechanism reminiscent of the operational modality of the Forget gate. Subsequent to this regulation, a vector is generated by employing the hyperbolic tangent (tanh) activation function. This function ensures that resultant values, which are influenced by h_{t-1} and x_t, are confined within the interval [-1, +1]. In the concluding step, an element-wise multiplication between the vector and the regulated values is conducted, thereby deriving the relevant information to be introduced to the cell state.

The equations for the input gate are [80]:

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i$$

$$\widehat{C_t} = \tanh(W_c.[h_{t-1}, x_t] + b_c$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widehat{C_t}$$

Where:

- ⊙ : Represents element-wise multiplication.

- tanh: Tanh activation function.

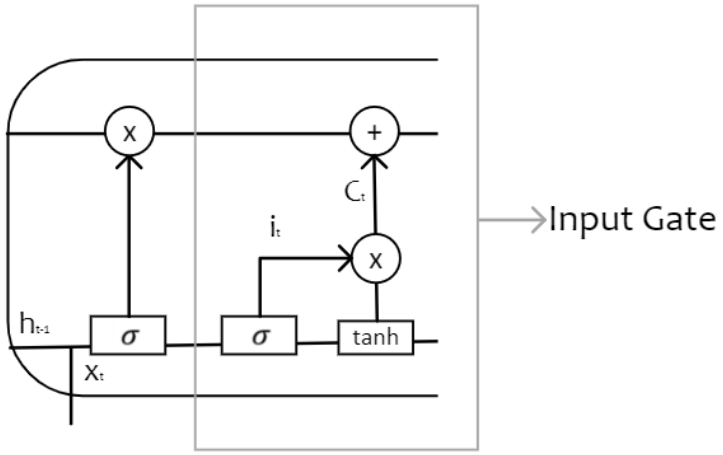

Figure 11. Input Gate

4.2.1.3. Output Gate:

In this terminal phase, the LSTM cell discerns which segments of information should be transmitted to the succeeding time instance, thereby relaying the contemporaneous knowledge accumulation.

The primary objective of the Output gate is to deduce the optimal lexical candidate to fill the lacuna.

$$o_t = \sigma(x_t * U_o + H_{t-1} * W_o)$$

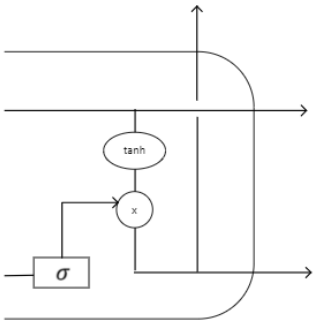

Figure 12. Output Gate

The operation underlying the output gate is fundamentally mathematical in nature. The equation governing the output gate, consistent with the architectural framework of its antecedent gates, guarantees that the resultant value is confined within the interval [0,1] owing to the employment of the sigmoid function. To determine the contemporaneous hidden state, the algorithm amalgamates the value o_t— representative of the output gate's value — with the hyperbolic tangent of the rejuvenated cell state, articulated as [81]:

$$H_t = o_t * \tanh(C_t)$$

In the subsequent stage, the token achieving the highest score in the output is identified as the prediction.

Delving into the mechanics, the responsibility of the output gate is to extract relevant information from the current cell state to be presented as the resultant output. Initially, the cell state is processed through the hyperbolic tangent function to produce a vector. This vector is then subject to modulation by the sigmoid function, which selects pertinent values primarily influenced by inputs h_{t-1} and x_t. The final operation involves combining the resultant vector values with the modulated values, preparing them for both output propagation and as input for the subsequent cell. The foundational equation governing the output gate is defined as:

$$o_t = \sigma(W_o \,.\, [h_{t-1}, x_t] + b_o)$$

This sophisticated integration of gates, complemented by mathematical computations, culminates in a comprehensive representation of the LSTM network, highlighting its prowess in data processing.

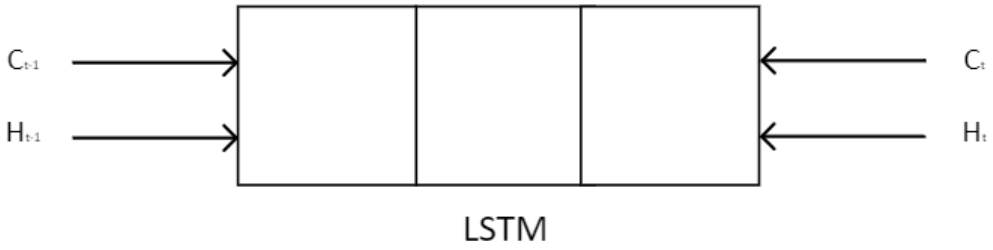

Figure 13. LSTM

This structured progression, encompassing the appraisal of prior data, amalgamation of the current input, and subsequent conveyance of the aggregated information, epitomizes a singular operational cycle of the LSTM, traditionally characterized as a single-time step.

The synergistic operation of these three gates, in conjunction with the memory cell, permits the LSTM unit to emulate the dynamics observed within a layer of neurons in a canonical feedforward neural network. Under this conceptualization, each LSTM "*neuron*" is endowed with both a latent and a present state, facilitating proficient processing of sequential data.

Within the domain of Recurrent Neural Networks (RNN), the Long Short-Term Memory (LSTM) represents a sophisticated architectural variant. Analogous to a rudimentary RNN, the LSTM upholds a latent state. In the framework, H_{t-1} signifies the latent state from the antecedent time instance, whereas H_t represents the latent state pertaining to the present time instance [82].

In a distinctive elaboration within the LSTM framework, an ancillary component, termed the cell state, is incorporated. This is denoted by C_{t-1} corresponding to the antecedent time instance and C_t for the contemporaneous instance. Within this context, the hidden state is conventionally characterized as the short-term memory, juxtaposed with the cell state which epitomizes the long-term memory. A graphical representation elucidating this bifurcated memory system can be observed in the accompanying illustration.

The LSTM network is a type of RNN which possesses the ability to learn and remember over long sequences and is less susceptible to the vanishing gradient problem as compared to the standard RNNs.

4.2.2. Pseudocode

The LSTM has a complicated cell structure that includes input, forget, and output gates. The pseudocode will provide an overview of the LSTM update process for a single time step:

```
Procedure LSTM_Cell(input_t, hidden_state_prev, cell_state_prev, parameters):
    # Extract parameters
    Wf, Wi, Wo, Wc, Uf, Ui, Uo, Uc, bf, bi, bo, bc = parameters
    # Concatenate hidden state with current input
    combined = Concatenate(hidden_state_prev, input_t)
    # Forget gate
    f_t = Sigmoid(Wf * combined + Uf * hidden_state_prev + bf)
       # Input gate
    i_t = Sigmoid(Wi * combined + Ui * hidden_state_prev + bi)
    C_tilda = Tanh(Wc * combined + Uc * hidden_state_prev + bc)
```

```
    # Update cell state
    cell_state_t = f_t * cell_state_prev + i_t * C_tilda
    # Output gate
    o_t = Sigmoid(Wo * combined + Uo * hidden_state_prev + bo)
    hidden_state_t = o_t * Tanh(cell_state_t)
    Return hidden_state_t, cell_state_t
Procedure LSTM_Network(input_sequence, initial_hidden_state, initial_cell_state, parameters):
    hidden_state = initial_hidden_state
    cell_state = initial_cell_state
    For each input_t in input_sequence:
        hidden_state, cell_state = LSTM_Cell(input_t, hidden_state, cell_state, parameters)
    EndFor
    Return hidden_state
```

In the LSTM pseudocode, the variables "W_f, W_i, W_o, W_c, U_f, U_i, U_o, U_c, b_f, b_i, b_o, b_c" represent the weights and biases for the various gates and cell state computations within the LSTM cell.

Forget Gate:

W_f: Weights for the forget gate associated with the current input.

U_f: Weights for the forget gate associated with the previous hidden state.

b_f: Bias term for the forget gate.

Input Gate:

W_i: Weights for the input gate associated with the current input.

U_i: Weights for the input gate associated with the previous hidden state.

b_i: Bias term for the input gate.

Output Gate:

W_o: Weights for the output gate associated with the current input.

U_o: Weights for the output gate associated with the previous hidden state.

b_o: Bias term for the output gate.

Cell State Update:

W_c: Weights for creating a new candidate cell state, associated with the current input.

U_c: Weights for creating a new candidate cell state, associated with the previous hidden state.

b_c: Bias term for the new candidate cell state.

In the LSTM cell:

The "W" matrices handle the input for each respective component.

The "U" matrices handle the previous hidden state for each respective component.

The "b" vectors are the biases for each respective component.

These parameters are learned during the training process to allow the LSTM to capture temporal dependencies in the data.

4.2.3. Mathematical Background

The Long Short-Term Memory (LSTM) model is a specific type of Recurrent Neural Network (RNN) architecture, introduced by Hochreiter and Schmidhuber in 1997, designed to alleviate the long-term dependency problem inherent in traditional RNNs.

Notations: Given a sequence $\{x_1, x_2, \dots, x_T\}$,at each time step t:

x_t is the input vector.

h_t is the hidden state vector.

c_t is the cell state vector.

The LSTM employs various gating mechanisms:

f_t denotes the forget gate's activation.

i_t denotes the input gate's activation.

o_t denotes the output gate's activation.

$\widetilde{c_t}$ represents a candidate cell state.

4.2.3.1. Equations:

The LSTM update equations are as follows:

$$f_t = \sigma\big(W_f . [h_{t-1}, x_t] + b_f\big)$$

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i)$$

$$\widetilde{c_t} = \tanh(W_c . [h_{t-1}, x_t] + b_c$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \widetilde{c_t}$$

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o$$

$$h_t = o_t \circ \tanh(c_t)$$

Where:

W_f ,W_i ,W_c , and W_o are weight matrices for the forget gate, input gate, cell update, and output gate respectively.

b_f ,b_i ,b_c , and b_o are bias vectors.

σ denotes the logistic sigmoid function.

tanh denotes the hyperbolic tangent function.

∘ represents element-wise multiplication.

4.2.3.2. Gating Mechanisms:

Forget Gate *(f_t):* Determines the proportion of the previous cell state c_{t-1} that should be retained. Values range between 0 (forget all) and 1 (retain all).

Input Gate *(i_t):* Controls the proportion of the new candidate cell state $\tilde{c}_t$ that should be added to the cell state.

Candidate Cell State *($\tilde{c}_t$):* Proposes a new cell state which is a combination of the current input and the previous hidden state.

Cell State *(c_t):* Updated by considering the forget gate's output, the previous cell state, and the contribution from the input gate and the candidate cell state.

Output Gate *(o_t):* Controls the proportion of the internal cell state c_t to expose to the external hidden state h_t.

4.3. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs), also termed as ConvNets, belong to a specialized subset of neural networks adept at processing data with a grid-like topology, primarily images. These networks have gained prominence in the domain of DL and play a pivotal role in the realm of computer vision. Within the broader scope of AI, computer

vision equips computing systems with the capability to analyze and interpret visual or image data.

Digital images, the primary input for CNNs, are binary codifications of visual stimuli. Such images are constituted by an array of pixels in a grid layout, where each pixel quantifies the color and luminosity to be manifested.

The proposition of CNNs can be credited to Yan LeCun in 1998. Their efficaciousness in image-related tasks, such as identifying a numeral present within an image, is noteworthy. The efficacy of CNNs in discerning patterns and tackling intricate tasks can be attributed to their architecture that draws parallels with cognitive functions of the human cerebrum [83].

To comprehend the intricacies of CNNs, a foundational understanding of the operational intricacies of neural networks is imperative. These networks emulate the human brain's capabilities in pattern discernment and problem resolution. Constitutively, a neural network comprises neurons stratified into multiple echelons: an input stratum, several intermediary (hidden) strata, and an output stratum. The intermediary strata's quantity often correlates with the problem's intricacy.

Data, upon ingress through these strata, facilitates pattern recognition by the neurons, culminating in the genesis of a representational construct termed a model. Post model training, the network employs it for prognostications on test datasets.

Prior to CNNs, image classification predominantly hinged on MLP. However, CNNs, as a refined iteration of ANN, exhibit enhanced efficacy, especially with matrix-structured data. In datasets like videos or images, intrinsic patterns are pivotal. Given CNNs' design to efficaciously extract such features, they emerge as the preferred choice for myriad applications necessitating matrix data pattern recognition.

Canonical Neural Networks are typified by three salient layers [84]:

1. ***Input Layer:*** This foundational layer receives the input data. The neuron count in this stratum matches the feature count of the proffered data. In imagery contexts, this is tantamount to the pixel count.

2. ***Hidden Layer:*** Succeeding the input layer, this stratum can be multifarious in a model depending on its architecture and data magnitude. Each hidden layer can encapsulate diverse neuron counts, usually surpassing the data's feature count. Data transference across layers involves a biphasic process: matrix multiplication of the preceding layer's output with the extant layer's mutable weights, succeeded by the addition of mutable biases. Subsequently, a nonlinear activation function is invoked.

3. ***Output Layer:*** Input from the terminal hidden layer is processed herein. A logistic function, like the sigmoid or softmax, refines the class-related output into probability metrics pertinent to each class.

The methodology of input data assimilation by the model and the sequential extraction of outputs through each layer is denominated as "*feedforward*". Post this, an error metric (e.g., cross-entropy or mean squared error) quantifies the model's deviation from expected outputs, serving as a performance index. The ensuing "*backpropagation*" phase employs derivative computations to attenuate this error, optimizing the model's precision.

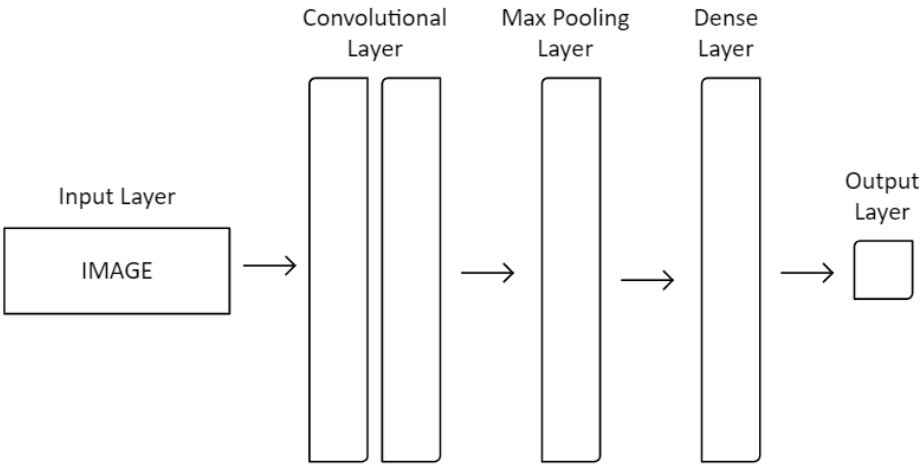

Figure 14. CNN Layers

4.3.1. CNN Architecture

A CNN, colloquially termed as covnet, encapsulates a sophisticated multi-layered architecture, each serving a distinct purpose in the comprehensive operation of the network. Delving into its structure, a prototypical CNN comprises the ensuing components:

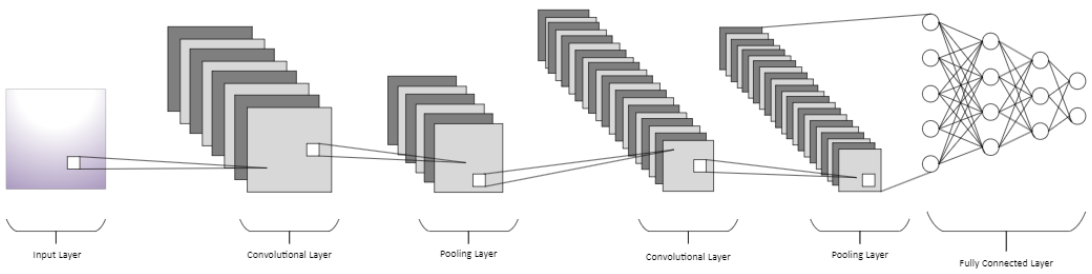

Figure 15. Layers of CNN

4.3.1.1. Input Layer

Primarily, this layer ingests raw image data introduced to the model. The spatial dimensions of the input are contingent upon the inherent attributes of the image, encapsulating its width, height, and depth.

4.3.1.2. Convolutional Layer

This pivotal layer is responsible for discerning features from the input matrix. It harnesses an ensemble of trainable filters or kernels—compact matrices of dimensions such as 2x2, 3x3, or 5x5—strategically superimposed on the input images. As these kernels traverse the image, they compute the dot product between their weights and the congruent image patch, culminating in the generation of feature maps. During the computation's forward propagation, every kernel navigates the entire span of the image's height and width, yielding a bi-dimensional output or an activation map for each kernel. The size of the step (stride) with which the kernel transits are modulated by the input image's dimensions. For an input of size $W \times W \times D$ and D_{out} number of kernels of spatial size F, with stride S and padding P, the resulting output volume dimensions can be computed using [85,86]:

$$W_{out} = \frac{W - F + 2P}{S} + 1$$

This formula engenders an output volume of size W_{out} x W_{out} x D_{out}.

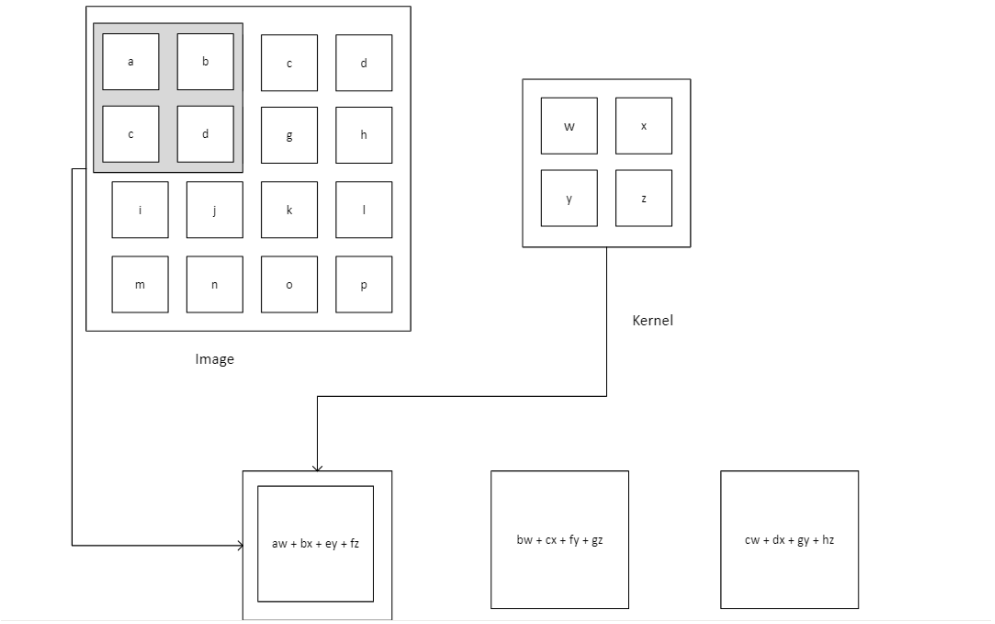

Figure 16. CNN Kernel

Crucially, convolutional layers integrate three paramount principles that have been instrumental in advancing computer vision research: sparse interactions, parameter sharing, and equivariant representations.

***Sparse Interaction**:* Contrary to the ubiquitous matrix multiplication observed in traditional neural network layers that delineates the interactions between every input-output pair, CNNs embrace sparse interactions. By designing the kernel to be more compact than the input, this modus operandi attenuates the requisite parameters, fostering enhanced computational efficiency and memory conservation.

***Parameter Sharing**:* A salient feature of CNNs is the universal application of a weight set for feature extraction across various spatial locales on the image. This paradigm operates on the premise that a particular feature's utility in one spatial domain suggests its potential applicability in others. Consequently, CNNs adopt shared parameters, wherein weights operationalized on a particular input section find repeated utility across the image.

***Equivariant Representation**:* Stemming from the principle of parameter sharing, CNNs manifest an attribute of equivariance to spatial translations. This intrinsic property denotes that specific transformations in the input induce corresponding shifts in the output."

In essence, CNNs, through their specialized architectural nuances, serve as formidable tools in extracting and recognizing intricate spatial patterns and hierarchies within image data.

4.3.1.3. Activation Layer

The activation layer is indispensable in infusing nonlinearity into the network's architecture. It applies an element-wise activation function to the outputs generated by the antecedent convolutional layer. Among the frequently employed activation functions are the ReLU, Tanh, and Leaky ReLU. Significantly, this layer preserves the spatial dimensions of its input, implying that an input volume of dimensions 32 x 32 x 12 would yield an output with identical dimensions.

4.3.1.4. Pooling Layer

Serving as an integral facet of convolutional neural networks, the pooling layer is designed to downsample the input representation, culminating in a marked reduction in computational exigencies. By diminishing the spatial dimensions of the input volume, the pooling layer not only expedites computational procedures but also optimizes memory utilization and mitigates the proclivity for overfitting. Incorporated intermittently within the network's architecture, this layer primarily harnesses two prevalent pooling techniques: max pooling and average pooling.

Within the realm of max pooling, the mechanism culls the maximal value from a designated neighborhood, while average pooling extracts the arithmetic mean of the values encompassed within a given neighborhood. To elucidate, the application of a max

pooling operation with filter dimensions of 2x2 and a stride of 2 would metamorphose an input volume of 32x32x12 into an output volume of 16x16x12.

At its core, the pooling layer operates by supplanting specific outputs in the network with a summary statistic derived from adjacent outputs. It assesses each slice of the input independently, engendering a diminution in the spatial dimensions of the representation, which, in turn, results in a concomitant reduction in computational overhead and requisite weights.

While an assortment of pooling methodologies exists—including the computation of the average over a rectangular region, the L2 norm over a similar region, or a weighted average anchored by the proximity to the central pixel—max pooling remains the preeminent choice in contemporary CNN architectures.

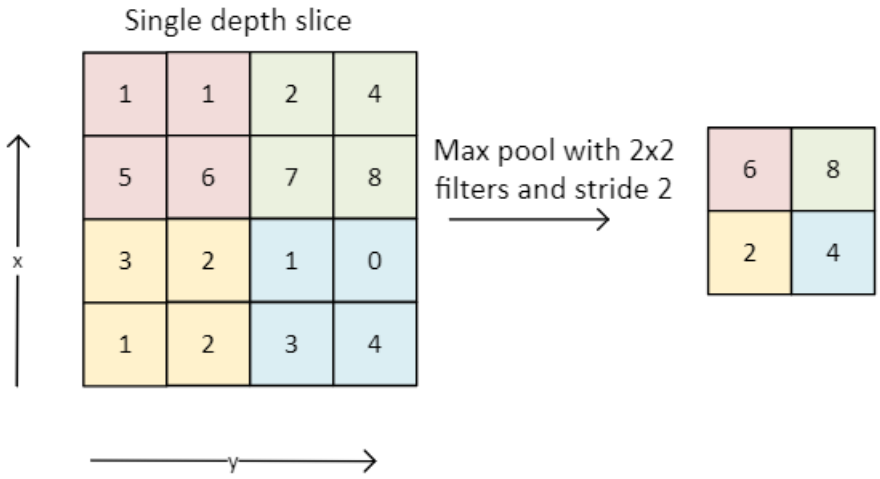

Figure 17. Max Pooling

The pooling layer, while central to achieving computational efficiency, also confers an element of translational invariance. This characteristic ensures that object recognition is robust to minor perturbations in its spatial positioning within the frame. Let the activation map dimensions be represented by *W x W x D*, the spatial extent of the pooling kernel by

F, and the stride by *S*. The dimensions of the resultant volume can be determined using the relationship:

$$W_{out} = \frac{W - F}{S} + 1$$

This relationship translates to an output volume with dimensions of W_{out} *x* W_{out} *x D.*

4.3.1.5. Flattening Layer

Sequential to the convolutional and pooling stages in the CNN schema, the feature maps undergo transformation in the flattening layer. In essence, this layer reconfigures the three-dimensional output from antecedent layers into a linear vector. This transmutation is imperative, paving the way for the ensuing feature vector to interface with the fully connected layer, wherein terminal, task-specific computations—such as classification or regression—are enacted.

4.3.1.6. Fully Connected Layer

This layer serves as the penultimate phase in the Convolutional Neural Network design. Acquiring input from the flattening stratum, it orchestrates the determinative classification or regression undertakings. Every neuron in this layer is intricately linked with all neurons both antecedent and subsequent to it, evoking the architectural paradigm endemic to traditional Fully Connected Neural Networks (FCNN). The computations intrinsic to this layer can be delineated as matrix multiplication endeavors succeeded by the annexation of a bias term. Essentially, this layer capitalizes on the features, meticulously extracted and refined by earlier layers, to pronounce the terminal inferences. At its core, the fully connected layer underpins the bijective mapping between the network's input and output representations.

4.3.1.7. Output Layer

Representing the terminal echelon of the CNN, the output layer receives the verdicts from the fully connected strata. Herein, a logistic apparatus, such as the sigmoid or softmax function, is invoked for classification-oriented tasks. These mechanisms morph the outputs for individual classes into probabilistic metrics, signifying the propensity of the input datum aligning with specific classes. Consequently, the output layer epitomizes the network's final prognostication by proffering a probability distribution across potential classes.

4.3.1.8. Non-Linearity Layers

Given that convolution fundamentally embodies a linear operation and the inherent nature of image data often manifests non-linear characteristics, it is customary to intercalate non-linearity within the activation maps. This is achieved by strategically positioning non-linearity layers subsequent to the convolutional strata.

Several variants of non-linear operations have been conceptualized, with the following being particularly prominent:

1. ***Sigmoid Function:*** Described by the expression

$$\sigma(K) = \frac{1}{1 + e^{-K}}$$

The sigmoid function compresses any continuous number into a domain spanning 0 to 1. However, it is imperative to note the sigmoid function's inherent limitation: during instances wherein the activation approximates either extremity of its domain, the gradient virtually becomes infinitesimal. This can inadvertently attenuate the gradient during the backpropagation process. Additionally, if a neuron persistently receives positive data inputs, the subsequent sigmoid function outputs maintain consistent polarities, engendering an oscillatory gradient update trajectory for the corresponding weight.

2. ***Tanh Function:*** The hyperbolic tangent function, or tanh, constrains a continuous number to a domain ranging from -1 to 1. Analogous to the sigmoid function, the tanh function can experience output saturation. However, it proffers the advantage of zero-centric outputs.

3. ***Rectified Linear Unit (ReLU):*** Over recent years, ReLU has garnered considerable attention and adoption. It is defined by the function $f(\kappa) = \max(0,\kappa)$, implying that activations with negative values are rectified to zero. Empirical studies suggest that ReLU often operates with enhanced efficiency, augmenting the convergence velocity by a factor of approximately six when juxtaposed against the sigmoid and tanh functions. Nonetheless, ReLU units exhibit a degree of fragility during the training process. They can succumb to "*death*" when subjected to substantial gradients, rendering them inert to subsequent updates. To obviate this pitfall, it is crucial to judiciously select an optimal learning rate.

4.3.2. CNN on Tabular Data

CNNs, distinguished primarily for their adeptness in image and video analyses, have recently been extrapolated to assess non-visual data categories, notably tabular data. This form of data, characteristically arrayed in rows and columns, is prevalent in sectors such as finance, healthcare, and e-commerce. Conventional methodologies for predictive analysis of tabular data have historically leveraged algorithms including decision trees, random forests, gradient boosting machines, and linear models. Nonetheless, contemporaneous research has underscored the potential of CNNs in binary classification paradigms utilizing tabular data, with one-dimensional convolution emerging as a quintessential approach.

Within this framework, the model conceptualizes each row as a sequential entity, utilizing filters of assorted dimensions to discern patterns across diverse feature groupings. The application of CNNs to tabular data is accompanied by multiple merits. Foremost, CNNs manifest proficiency in ascertaining hierarchical feature representations directly from

unprocessed data, negating the imperative for manual feature engineering. Subsequently, CNNs facilitate end-to-end learning, whereby raw input is directly harnessed for predictive purposes, streamlining the training procedure. Moreover, the integration of multifarious filters capacitates the model to assimilate and distinguish a plethora of patterns intrinsic to the data. Conclusively, given judicious architectural and hyperparametric selections, CNNs can equate or occasionally outperform the efficacies of traditional machine learning paradigms [87].

Nevertheless, employing CNNs for tabular data introduces specific challenges. Notably, tabular data is devoid of the intrinsic spatial or temporal relationships that are emblematic of image or sequential data, potentially engendering imprecise model interpretations. Additionally, CNNs may exhibit deficiencies in handling sparse data, a recurrent characteristic of tabular configurations. The disparate magnitudes inherent to varied features can complicate convolution implementations, necessitating meticulous preprocessing. Furthermore, analogous to numerous deep learning architectures, CNNs predominantly operate in an opaque manner, obscuring the interpretability of their predictions.

The foundational components of a CNN, encompassing the convolutional, pooling, and fully connected layers, adhere to distinct operational principles when interfacing with tabular data. Within convolutional strata, pivotal to CNNs, an array of filters is deployed on the input. For tabular data, it's typically transformed into a 3D configuration (samples, features, 1) to align with the model's requirements. These echelons interpret each row (or instance) as a sequential datum, implementing filters across this continuum to abstract pertinent features. For illustration, X denotes a row from the tabular dataset with length n, and let F symbolize a filter or kernel with length m adopted by the convolutional stratum.

The convolution of X and F is articulated as:

$$Z(i) = \sum_{j=0}^{m-1} F(j) * X(i+j)$$

In this expression, *i* spans from 0 to $n-m+1$, and *Z* represents the feature map engendered by the convolutional layer. This operation entails element-wise multiplication of the coinciding segments of *X* and *F*, succeeded by an aggregation. This computation is iteratively applied to all rows (instances) in the tabular dataset.

In the domain of CNNs, it is customary for individual convolutional layers to harbor an array of filters. Each filter is uniquely attuned to discern specific patterns within the input dataset. Upon undergoing convolution operations, the derived data is subsequently propagated through a non-linear activation function, with the ReLU being a prevalent choice. Each filter or kernel within the convolutional layer can be construed as an ensemble of weights, tailored during the model's training phase to identify and respond to discrete patterns within the input.

By systematically sweeping across the input, these filters execute the convolution operation, engendering a feature map. This map accentuates regions where the input manifests patterns resonating with the filter's configuration. Given the innate ability of each filter to specialize in discerning distinct patterns, the overarching CNN model is equipped to recognize a diverse array of relationships, encompassing potential complexities inherent to the data. This intrinsic capability for automated feature extraction is a salient merit underscoring the deployment of CNNs for tabular data analysis.

Sequential to the convolutional layers within the conventional CNN architecture reside the pooling layers. These are meticulously crafted to curtail the spatial dimensions—explicitly, the height and width—of the processed data. This dimensionality reduction not only mitigates computational overhead but also serves as a prophylactic against

overfitting. In the milieu of tabular data, this dimensionality reduction parallels a truncation in the quantum of features conveyed to the ensuing layers.

For instance, if one denotes the feature map elicited from a convolutional layer as *Z*, and postulates the employment of max pooling with a designated pool size *p*, the ensuing pooled feature map *P* can be formalized as:

$$P(i) = \max(Z[i, i+1, \ldots, i+p-1)]$$

Herein, i oscillates between 0 and the quotient of the length of *Z* and *p*. Each constituent of *P(i)* encapsulates the zenith value within a window of dimension *p* in *Z*.

When delineating tabular data, the term "*spatial dimension*" is synonymous with the dataset's feature count. Among the myriad pooling techniques, max pooling and average pooling are predominant. While the former retains the apical value from each segment of the feature map, the latter consolidates and retains the mean value. Through the preservation of pivotal (maximal) or emblematic (average) values, pooling strata effectuate a non-linear form of data compression.

Situated typically towards the terminal portion of the architecture are the fully connected layers, which are pivotal in effectuating intricate reasoning based on the distilled features from prior convolutional and pooling strata. Each neuron within these layers is interlinked with its antecedent counterparts, enabling the synthesis of non-linear feature amalgamations.

For binary classification paradigms, the terminal fully connected layer is typically equipped with a solitary neuron, which dispenses the prediction metric. This metric is then channeled through a sigmoid activation function, calibrating the output to lie within

the [0, 1] continuum, thereby conveying the probability estimate of the positive class designation.

Fully connected strata serve as the neural epicenters where intricate reasoning materializes. From a mathematical vantage point, each neuron in this layer computes a weighted summation of its inputs, augments this with a bias metric, and subsequently channels this aggregate through an activation function.

Consider the scenario wherein Y symbolizes the linearized output emanating from the antecedent pooling layer, possessing a length n. W be the representative weight vector, and let b stand for the bias coefficient pertinent to a neuron embedded within the fully connected layer. Under such stipulations, the output O furnished by the neuron can be delineated by the equation:

$$O = F(\sum_{i=1}^{n} W(i)Y(i) + b)$$

In this context, the activation function, represented as F, commonly adopts the ReLU for intermediary layers. However, when confronted with a binary classification undertaking, the output layer conventionally employs the sigmoid function.

For binary classification architectures within the CNN paradigm, the concluding layer is typified by the presence of a singular neuron. The output generated by this neuron is construed as the probability of an instance being affiliated with the positive class. When applied to tabular data, this modus operandi empowers the model to render a binary verdict contingent upon the weighted significance of an assortment of discerned features present in the dataset, thereby culminating in the final classification verdict.

Consequently, the holistic operation of a neuron nestled within the fully connected layer, assimilating the activation function into the calculus, can be succinctly encapsulated by the equation:

$$O = f(W.Y + b)$$

Wherein, O is the neuron's output, W signifies the weight vector, Y embodies the linearized output derived from the preceding pooling layer, and b stands for the bias coefficient.

4.3.3. Pseudocode

```
Procedure CNN_Training(data, labels, epochs, learning_rate):
    Initialize convolutional filters F, weights W, and biases b randomly
    For epoch in 1 to epochs:
        For each (image, target) in (data, labels):
            output = forward_pass(image)
            loss = compute_loss(output, target)
            gradients = backpropagate_loss(loss)
            update_parameters(gradients, learning_rate)
        EndFor
    EndFor

Procedure forward_pass(image):
    // Convolution Layers
    For each convolution_layer in convolution_layers:
        convolved_feature = convolution(image, F) + b
```

```
        activated_feature = activation_function(convolved_feature)
        pooled_feature = pooling(activated_feature)
        data = pooled_feature  // Setting output as input for next layer
    EndFor

    // Fully Connected Layers
    flattened = flatten(pooled_feature)
    For each fc_layer in fully_connected_layers:
        z = W * flattened + b
        activated = activation_function(z)
        flattened = activated
    EndFor

    return activated

Procedure compute_loss(output, target):
    return loss_function(output, target)

Procedure backpropagate_loss(loss):
    // Compute gradients for fully connected layers
    For each fc_layer in reversed(fully_connected_layers):
        delta = loss_derivative(output, target) * activation_derivative(z)
        gradient_W = delta * output_previous_layer.transpose()
        gradient_b = delta
```

```
        loss = W.transpose() * delta  // Backpropagate the loss to the previous layer
    EndFor

    // Compute gradients for convolutional layers
    For each conv_layer in reversed(convolution_layers):
        delta = ...  // Gradient from pooling and activation function
        gradient_F = convolution(input_previous_layer, delta)
        gradient_b = delta
    EndFor

    return gradients for all F, W, and b

Procedure update_parameters(gradients, learning_rate):
    For each parameter in (F, W, b):
        parameter = parameter - learning_rate * gradients[parameter]
    EndFor
```

4.3.4. Mathematical Background

1. Convolution Layer:

The foundational component of CNNs is the convolution operation. Given an input feature map

I and a filter (or kernel) K, the convolution operation is defined as [88]:

$$(I * K)(x, y) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} I(i, j).K(x - i, y - j)$$

In practice, for a 2D input with dimensions $W \times H$ and a filter of dimensions $F_W \times F_H$, the equation simplifies to:

$$(I * K)(x, y) = \sum_{i=1}^{F_W} \sum_{j=1}^{F_H} I(x + i - 1, y + j - 1).K(i, j)$$

2. Pooling Layer:

Pooling layers are employed to reduce the spatial dimensions of the feature maps, leading to decreased computational demands. The most common form is max pooling. Given a region R in the feature map, *max* pooling is:

$$maxpool(R) = max_{(x,y)\in R} R(x, y)$$

Other pooling strategies include average pooling and min pooling.

3. Fully Connected Layer:

A fully connected layer in a CNN operates similarly to traditional neural network layers. If X represents the flattened output from the previous layer (or input) with dimension $N \times 1$, and W is the weight matrix of dimension $M \times N$ (where M is the number of neurons in the current layer), the output Y is:

$$Y = W.X + b$$

Here, b is the bias vector of dimension $M \times 1$.

4. Activation Functions:

After each convolutional or fully connected layer, an activation function is applied element-wise to introduce non-linearity. Common functions include:

$$ReLU\ (Rectified\ Linear\ Unit): f(x) = \max(0, x)$$

$$Sigmoid: f(x) = \frac{1}{1 + e^{-x}}$$

$$Tanh: f(x) = \tanh(x)$$

5. Batch Normalization:

Batch normalization often follows the convolution operation, stabilizing the activations of the neurons. Given a set of activations X, the normalized activation $\hat{X}$ is:

$$\widehat{x_i} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

Where μ is the batch mean, σ^2 is the batch variance, and ϵ is a small constant for numerical stability.

6. Loss Functions:

Depending on the task, different loss functions are employed. For classification tasks, the categorical cross-entropy loss is commonly used:

$$L = -\sum_{i=1}^{C} y_i \log(\widehat{y_i})$$

Where *C* is the number of classes, y_i is the true label, and $\hat{y}_i$ is the predicted probability.

4.4. Random Forest (RF)

Random Forest (RF), a supervised learning algorithm extensively employed in diverse sectors, is noteworthy for its robust performance and simplicity. Originated and trademarked by Leo Breiman and Adele Cutler, the foundational philosophy of this algorithm is rooted in the principle of ensemble learning. This principle posits that the amalgamation of multiple learning models amplifies the overall effectiveness of predictive outcomes. Within the context of RF, this ensemble comprises numerous decision trees [89].

These individual decision trees are cultivated on assorted subsets of the given dataset, a practice generally affiliated with the bagging method. The objective of this approach is to augment the predictive accuracy and robustness of the algorithm, culminating in a reliable mechanism proficient in addressing both regression and classification challenges efficaciously.

The functionality of the RF algorithm is strikingly straightforward. When assigned a classification task, it constructs a multitude of decision trees on a variety of samples and subsequently establishes the final prediction based on a majority vote culled from these trees. The term "*forest*" in RF thus symbolizes these various decision trees, with the number of trees directly influencing the algorithm's precision and problem-solving prowess. A more abundant "forest" of trees correlates with superior predictive capabilities of the model.

RF exhibits remarkable flexibility, delivering commendable results even in the absence of meticulous hyperparameter tuning. Its versatility and user-friendly nature have facilitated its broad-spectrum application in numerous fields, ranging from healthcare to finance. The algorithm's capability to effectively manage intricate problems and consistently reliable performance further contribute to its popularity. As a result, the RF

algorithm is arguably one of the most recurrently implemented machine learning algorithms, owing to these unique advantages.

4.4.1. Ensemble Methods

RF, a method within supervised learning algorithms, fabricates an array of decision trees, amalgamating their output to procure a prediction that is both more precise and consistent. The versatility of this method is profound, as it effectively addresses both classification and regression tasks, which constitute a significant majority of the machine learning systems currently in operation. The hyperparameters employed in RF closely resemble those used within a decision tree or a bagging classifier, thereby obviating the necessity of integrating a decision tree with a bagging classifier, courtesy of the user-friendliness inherent in the classifier-class of RF.

Within the sphere of ensemble learning, rather than depending on a solitary predictive model, an array of machine learning models, referred to as weak learners, are developed. The amalgamation of these weak learners results in a strong learner, which dispenses comprehensive predictions across all targeted classes with considerable precision. Ensemble learning methodologies encompass an assortment of classifiers, for instance, decision trees, and the predictions procured from these are aggregated to ascertain the most frequently occurring outcome.

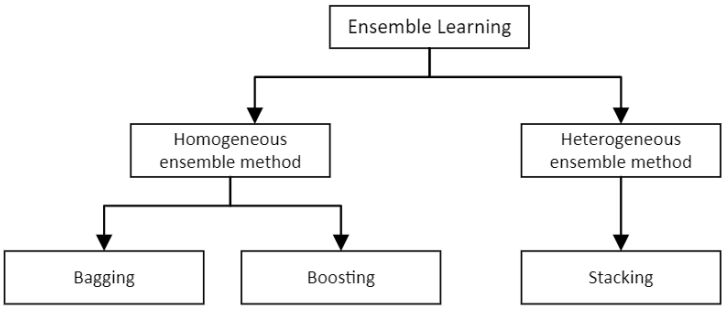

Figure 18. Ensemble Learning

Bagging and boosting stand as noteworthy instances of ensemble techniques. The bagging procedure, pioneered by Leo Breiman in 1996, involves selecting a random assortment of data from a training set with replacement. This allows individual data points to be selected repeatedly. Upon generating numerous data samples, each is trained independently. Depending on the nature of the task at hand - whether it is regression or classification - the aggregate or majority of the subsequent predictions offer a more accurate approximation. This methodology is predominantly employed to mitigate variance within datasets characterized by a high degree of noise.

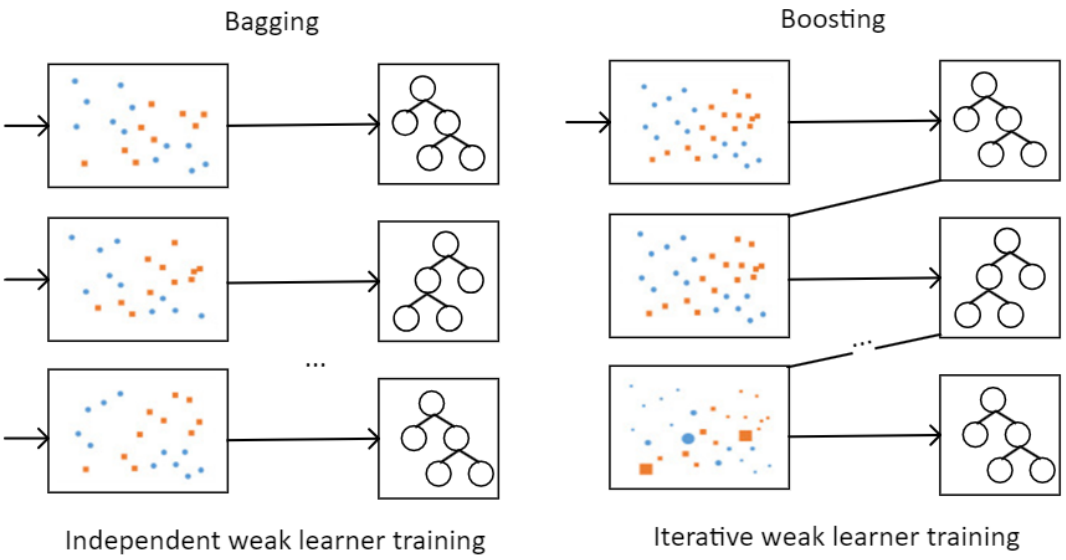

Figure 19. Bagging and Boosting

RF implements the bagging method, otherwise known as bootstrap aggregation. This process commences with an arbitrary selection of original data, which is subsequently arranged into discrete samples termed as "*Bootstrap Samples*" within a procedure denoted as "*Bootstrapping*". Subsequently, the models are trained independently, yielding a plethora of outcomes, a step classified as "*Aggregation*". In the terminal stage, the diverse results are amalgamated, and the resultant output is predicated upon majority voting. This procedure, recognized as "*Bagging*", is executed utilizing an Ensemble Classifier [90].

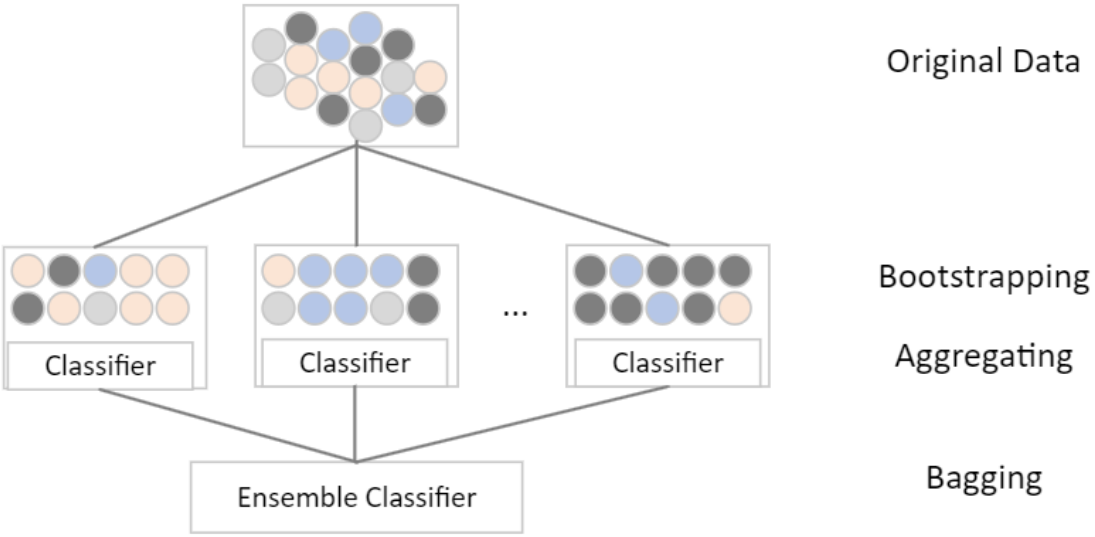

Figure 20. Ensemble Classifier

The operational mechanics of bootstrapping within the RF algorithm encompass both row and feature sampling with replacement preceding the training phase of the model. Given that the sampling procedure is performed with replacement, a proportion equivalent to nearly one-third of the data is not exploited during the model's training. This subset of data is referred to as out-of-the-bag samples. Performing an evaluation of the model utilizing these out-of-the-bag samples provides insights into its prospective performance on the test dataset.

4.4.2. Decision Tree vs Random Forest

The RF algorithm, an aggregation of multiple decision trees, serves as a crucial element within supervised machine learning. An initial comprehension of decision trees proves beneficial when exploring the intricacies of the RF algorithm. A decision tree embarks on its path with a fundamental question. Additional queries are subsequently established to reach a definitive conclusion, each signifying a decision node within the tree employed to bifurcate the data. Following either the affirmative or negative branch culminates in

the ultimate decision, represented by the leaf node. The evaluation of the split's quality is performed using various metrics such as Gini impurity, information gain, or MSE.

Decision trees employ a diagrammatic tree structure akin to a flowchart to elucidate predictions derived from a sequence of splits based on distinct features, initiating from a root node and culminating in a conclusion ascertained by leaf nodes. A decision tree consists of three primary elements: a root node, decision nodes, and leaf nodes. The root node initiates the segmentation of the population. The resulting nodes post-separation of a root node are termed decision nodes. A node where further bifurcation is infeasible is referred to as a leaf node. The procedure for selecting the root node is contingent on a method of prioritizing features.

Notwithstanding their widespread application in supervised learning algorithms, decision trees may be susceptible to certain issues, notably bias and overfitting. However, the amalgamation of multiple decision trees into a collective ensemble, as exemplified in the RF algorithm, augments the accuracy of results, particularly when the constituting trees are uncorrelated. RF, a variant of the bagging method, generates multiple decision trees, each predicated on distinct subsets of the original dataset, thereby efficaciously circumventing the issue of overfitting. This algorithm possesses the capacity to tackle both classification and regression problems, further enhancing its versatility.

A salient distinction between decision trees and RF is observable in the procedure of formulating rules. In the context of a training dataset inclusive of features and labels, a decision tree will devise a set of stipulations that underpin its predictive output.

"*Deep*" decision trees, whilst effective in certain contexts, can be prone to overfitting. Contrarily, RF often rectifies this by randomly selecting subsets of features to generate smaller trees, subsequently aggregating them into subtrees. This technique may not always yield successful results, and it has the potential to slow computational processes

contingent on the quantity of trees constructed. However, it remains a robust method for diminishing overfitting.

Compared to this, a single decision tree offers a more expedient computational process, but could potentially incur the risk of overfitting should it be allowed to proliferate to its maximum depth. Inversely, RF mitigates overfitting by deriving forests from subsets of data, with the final output contingent on either an average or a majority rating. This eradicates the dependency on any singular set of formulae, further illustrating the flexibility and robustness of the RF approach [91].

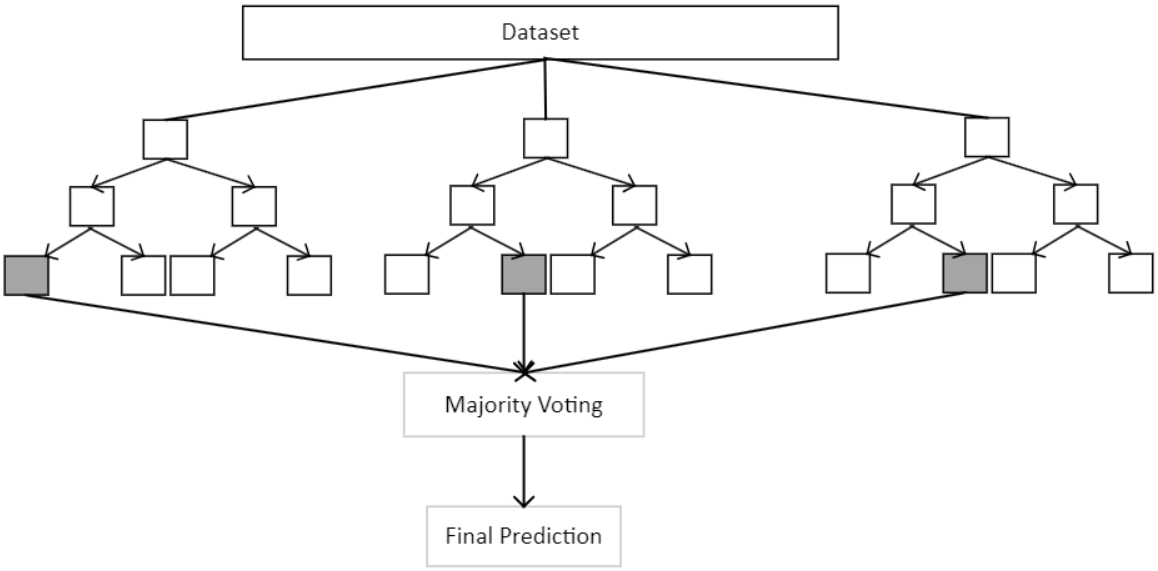

Figure 21. RF Classification

4.4.3. Pseudocode

Define the number of trees in the forest

number_of_trees = n

```
# Define the maximum depth of the trees
max_depth = d

# Initialize an empty list to hold the forest
forest = []

# For each tree in the forest
for i in range(number_of_trees):
    # Create a bootstrap sample of the data
    bootstrap_sample = create_bootstrap_sample(data)

    # Grow a decision tree on the bootstrap sample
    tree = grow_decision_tree(bootstrap_sample, max_depth)

    # Add the tree to the forest
    forest.append(tree)

# Define a function to make predictions with the forest
function predict(forest, new_data):
    # Initialize an empty list to hold the predictions of the trees
    tree_predictions = []

    # For each tree in the forest
    for tree in forest:
```

```
        # Make a prediction with the tree
        prediction = tree.predict(new_data)

        # Add the prediction to the list of predictions
        tree_predictions.append(prediction)

    # Return the most common prediction among the trees
    return most_common(tree_predictions)
```

4.4.4. Mathematical Background

The RF algorithm is a robust ensemble learning method. Its mathematical background lies in the principles of decision tree learning, bootstrapping, and averaging.

4.4.4.1. Decision Tree Learning

The fundamental building block of a Random Forest is the decision tree, specifically the Classification and Regression Tree (CART). A decision tree partitions the feature space recursively. At each internal node of the tree, a decision is made based on a feature threshold, directing the data to the left or right child node. This process continues until terminal nodes (leaves) are reached. The goal is to structure the tree such that data instances within each leaf node are as homogeneous as possible with respect to the target variable. The decision criteria can be quantified using measures such as Gini impurity or entropy for classification and variance for regression.

4.4.4.2. Bootstrapping

The RF introduces randomness into the model training process to ensure diversity among the trees. This is achieved through bootstrapping, a resampling technique. For each tree

to be trained, a bootstrap sample (a random sample with replacement) of the training data is drawn. This sample serves as the training data for that particular tree.

4.4.4.3. Feature Randomness

Another source of randomness is introduced during the node split decision. Instead of evaluating all features to decide the best split, a random subset of features is selected, and the best split feature is chosen from this subset. This practice further decorrelates the trees, enhancing the forest's generalization ability.

4.4.4.4. Aggregation

Once the forest is constructed, predictions are made by aggregating the predictions of all trees. For regression, this is typically the average prediction, while for classification, it is the majority vote.

Given a training dataset $D = \{[x_1, y_1), \dots, (x_n, y_n)$ where $x_i$the input features and Y_i are are the output labels or values:

1. For b=1 to B (where B is the number of trees in the forest):
    - Draw a bootstrap sample D $*$ of size N from D.
    - Grow a decision tree T_b on D $*$. At each node:
        - Select m features at random from the full set of features.
        - Split the node using the best split among the m features.
        - Continue until a stopping criterion is met.
2. For prediction with a new sample x:

- For regression:

$$\hat{y}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

- For classification:

$$\hat{y}(x) = majority\ \{T_1(x), T_2(x), \dots . . T_B(x)\}$$

The strength of the RF algorithm lies in its capacity to reduce variance by averaging multiple deep decision trees, each of which could have high variance and low bias on its own. The combination results in a model that retains the low bias of individual trees while significantly reducing variance.

4.5. eXtreme Gradient Boosting (XGBoost)

XGBoost, an acronym for eXtreme Gradient Boosting, represents an open-source embodiment of the gradient-boosted trees algorithm, encapsulating an ensemble learning methodology. The algorithm, positioned within the realm of robust and supervised machine learning, is extensively employed across classification and regression tasks. Its increased utilization, notably discernible in Kaggle competitions, is attributable to its unparalleled predictive accuracy and simplicity of operation.

The algorithm's design prioritizes speed, efficacy, and overall performance, making it particularly adept at processing large datasets. XGBoost's primary advantage is its relative independence from extensive parameter optimization or tuning it demonstrates commendable functionality immediately post-installation. It capitalizes on the power of parallel tree boosting, thereby enhancing efficiency and establishing itself as a preferred option for machine learning tasks encompassing regression, classification, and ranking problems.

XGBoost's potency is derived from its innovative amalgamation of several machine learning concepts and algorithms, including supervised machine learning, decision trees, ensemble learning, and gradient boosting. In the realm of supervised machine learning, algorithms are employed to train a model using a dataset with predefined labels and features. This trained model is subsequently utilized to predict the labels of unseen datasets.

By coordinating the predictions of multiple weaker models, XGBoost succeeds in generating a more robust and dependable prediction. Its capacity to manage large datasets

and consistently deliver cutting-edge performance accords it a privileged status among machine learning practitioners. Additionally, XGBoost furnishes a scalable, distributed approach for training gradient-boosted decision tree models, thereby consolidating its position as a leading machine learning library.

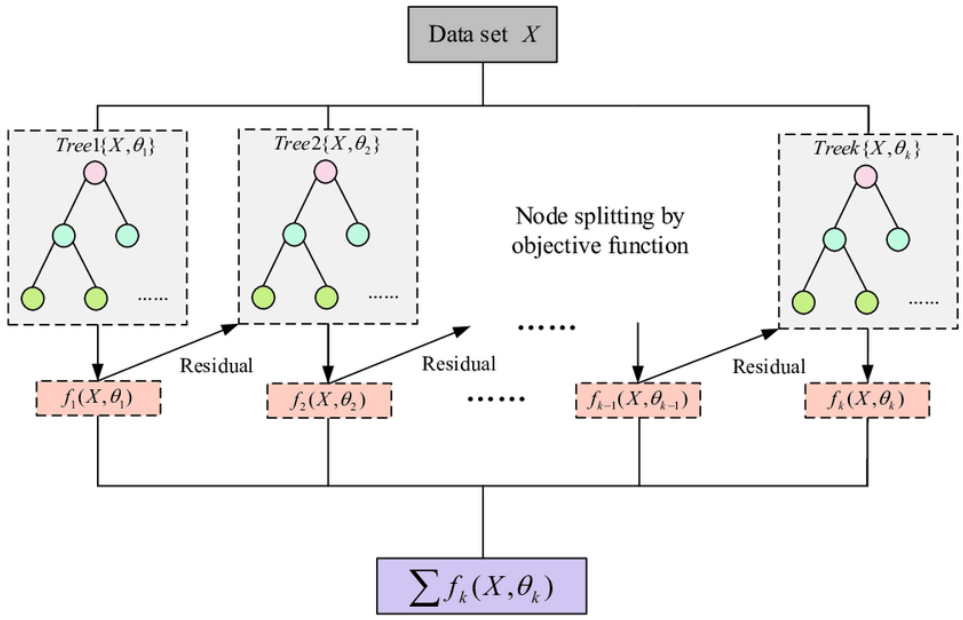

Figure 22. Gradient-Boosted Decision Tree

GBDT epitomize an ensemble learning algorithm, analogous to a random forest method, used for classification and regression applications. Ensemble learning embodies the synthesis of multiple machine learning algorithms to cultivate a superior predictive model. Both GBDT and RF engender a model comprising an array of decision trees. The critical divergence, however, arises in the methodologies employed to construct and amalgamate these trees.

The term "*gradient boosting*" derives from the concept of "*boosting*", which seeks to strengthen a solitary, feeble model by incorporating it with numerous other weak models, thereby engendering a robust predictive model. This process, when extended to gradient boosting, is formalized as a gradient descent algorithm applied over an objective function. In this paradigm, gradient boosting delineates targeted outcomes for the subsequent model in a bid to minimize error. These targeted outcomes are rooted in the gradient of the error for each case, giving rise to the term 'gradient boosting'.

GBDTs espouse an iterative methodology to train an ensemble of shallow decision trees, employing the error residuals from the preceding model to fit the succeeding one. The ultimate prediction manifests as a weighted aggregate of all predictions proffered by the trees. While the technique of "*bagging*" utilized in random forests seeks to curtail variance and overfitting, "*boosting*" in GBDT concentrates on diminishing bias and under fitting.

XGBoost is a highly accurate and scalable realization of gradient boosting that optimizes computational capabilities for boosted tree algorithms. Its main objectives are augmenting machine learning model performance and expediting computational speed. Unlike GBDT, wherein trees are constructed sequentially, XGBoost fabricates trees in parallel, using a level-wise strategy. It scrutinizes gradient values and harnesses these partial sums to assess the quality of splits at all possible points within the training set.

XGBoost's strength resides in its scalability, enabling swift learning via parallel and distributed computing while optimizing memory usage. Ensemble learning methodologies, such as XGBoost, offer a mechanism to amalgamate the predictive capacities of multiple learners, yielding a single model that aggregates output from several models. The base learners forming the ensemble could be derived from a single learning algorithm or multiple different ones. Bagging and boosting represent two commonly utilized ensemble learning techniques, typically implemented with decision trees [91].

Bagging seeks to reduce the variance in any learner by using multiple decision trees, generated concurrently. Training these learners entails using data sampled with replacement, with the final prediction being the averaged output from all the learners.

In contrast, boosting builds trees in a sequential manner, with each successive tree striving to minimize the errors of its predecessor. Each tree learns from preceding trees and

updates the residual errors, enabling the subsequent tree to learn from an updated version of the residuals. The base learners in boosting are weak learners characterized by high bias and a predictive power just marginally superior to random guessing. By effectively combining these weak learners, boosting generates a strong learner that significantly curtails both bias and variance.

In contrast to bagging techniques such as RF, where trees are grown to their maximum extent, boosting employs trees with fewer splits, resulting in highly interpretable small trees. Optimal determination of parameters such as the number of trees or iterations, the learning rate for gradient boosting, and the depth of the tree, can be achieved through validation techniques like k-fold cross-validation. Overfitting can be a risk with a large number of trees, hence determining the stopping criteria for boosting necessitates careful consideration.

The ensemble methodology inherent in gradient boosting encompasses three fundamental phases. Initially, a model F_0 is delineated to predict the target variable y, which is then connected to a residual, determined as (y - F_0). Subsequently, a novel model h_1 is fitted to the residuals stemming from F_0. Ultimately, F_0 and h_1 are amalgamated to engender *F1*, a boosted version of F_0. In this way, the mean squared error derived from F_1 is less than that obtained from F_0:

$$F_1(X) < -F_0(X) + h_1(X)$$

The performance can be further optimized by modeling residuals from F_1 and creating a subsequent model, F_2:

$$F_2(X) < -F_1(X) + h_2(X)$$

This process is reiterated for "*m*" iterations, until the residuals are minimized to the greatest extent possible:

In this scenario, the appended learners do not disrupt the functions established previously. Rather, they provide distinctive information to aid in error reduction:

$$F_m(X) = F_{(m-1)}(X) + h_m(X)$$

A salient attribute of XGBoost is its adeptness in managing missing values, rendering it an appropriate choice for handling empirical data with missing values without the need for extensive pre-processing. Additionally, XGBoost supports parallel processing, which expedites model training on extensive datasets within a reasonable timeframe.

$$F_m(X) < -F_{m-1}(X) + h_m(X)$$

XGBoost finds versatile applications, encompassing, but not limited to, Kaggle competitions, recommendation systems, and click-through rate predictions. An additional advantage of XGBoost lies in its extensive customizability, permitting meticulous tuning of myriad model parameters for optimized performance.

4.5.1. Mathematical Background

The underpinning mathematics of XGBoost involves the amalgamation of individual decision tree predictions. Each decision tree provides prediction scores, with the cumulative output of all the trees forming the ultimate prediction. A crucial aspect is the compensatory behavior exhibited by the trees; each tree endeavors to correct the weaknesses or errors of its predecessor, optimizing the total prediction.

The ensemble of decision trees can be encapsulated in a mathematical expression. According to this expression, each tree contributes to the final prediction, and can be viewed as the summation of individual tree functions. Every tree, denoted by a function, ingests the input data and produces an output, the aggregation of which yields the final predicted value:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad f_k \in F$$

Here, *"K"* is the number of trees, *"f"* is the functional space of *"F"*, and *"F"* is the set of possible CARTs. This equation captures the essence of the model and the interplay of the individual decision trees within it.

The objective function for the aforementioned model is provided as:

$$\hat{y}_i = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Here, the first term represents the loss function, and the second term constitutes the regularization parameter. Instead of endeavoring to learn the entire decision tree in a single step, which could complicate the optimization process, an additive strategy is employed. This strategy involves minimizing the loss from previous learning stages and integrating a new tree to enhance the predictive model, as outlined above.

$$\hat{y}_i = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = f_1(x_i) = f_1(x_i) + \hat{y}_i^{(0)}$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = f_2(x_i) + \hat{y}_i^{(1)}$$

….

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) \; = f_t(x_i) + \; \hat{y}_i^{(t-1)}$$

The objective function for the aforementioned model can be delineated as follows:

$$obj(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i)$$

$$= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)}) + f_t(x_i) + \; \Omega(f_t) + constant$$

$$obj(t) = \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \sum_{i=1}^{t} \Omega(f_i)$$

$$= \sum_{i=1}^{n} [2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + \; f_t(x_i)^2] + \; \Omega(f_t) + \text{constant}$$

Incorporating the second-order Taylor series expansion into this analysis:

$$obj(t) = \sum_{i=1}^{n} [l(y_i - (\hat{y}_i^{(t-1)}) + \; g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \; \Omega(f_t) + constant$$

where g_i and h_i can be defined as:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i - (\hat{y}_i^{(t-1)})$$

$$h_i = \partial^2{}_{\hat{y}_i^{(t-1)}} l(y_i - (\hat{y}_i^{(t-1)})$$

Simplifying and removing the constant:

$$\sum_{i=1}^{n} [\, g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \; \Omega(f_t)$$

Prior to exploring the details of the regularization term, it's essential to first define the framework of the model:

$$f_t(x) = \omega_{q(x)}\,, \omega \in R^T\,, q: R^d \to \{1,2,\ldots.,T\}$$

In this scenario, 'w' represents the vector of scores located on the leaves of the tree, 'q' is a function that allocates each data point to its corresponding leaf, and 'T' denotes the total number of leaves. Subsequently, the regularization term can be articulated as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

The objective function becomes:

$$obj(t) \approx \sum_{i=1}^{n} [g_i \omega_{q(x_i)} + \frac{1}{2} h_i {\omega^2}_{q(x_i)}] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

The aforementioned expression is then simplified:

$$obj(t) = \sum_{j=1}^{T} [G_j \omega_j + \frac{1}{2}(H_j + \lambda)\, \omega_j^2] + \gamma T$$

where,

$$G_j = \sum_{i \in I_j} g_i$$

$$H_j = \sum_{i \in I_j} h_i$$

In the specified equation, each component of w_j is independent. The optimal outcome for a predefined structure, q(x), and the maximum achievable reduction in the objective can be calculated with the following expression:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}$$

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G^2{}_j}{H_j + \lambda} + \gamma T$$

The parameter gamma in the above expression represents the pruning parameter, that is, the minimum information gain necessary to perform a split.

The quality of a tree is assessed by optimizing its structure one level at a time, rather than optimizing it holistically. This procedure specifically involves partitioning a leaf into two distinct leaves and assessing the subsequent improvement in the score that is derived from this action.

$$Gain = \frac{1}{2}\left[\frac{G^2{}_L}{H_L + \lambda} + \frac{G^2{}_R}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma$$

XGBoost, an eminent implementation of gradient boosting, encompasses an array of distinct features:

Regularization: XGBoost introduces both L1 and L2 regularization to incur penalties on intricate models, which assists in circumventing overfitting.

Handling of Sparse Data: XGBoost can proficiently manage absent values or data sparsified due to processes such as one-hot encoding, courtesy of its sparsity-aware split-finding algorithm.

Weighted Quantile Sketch: In contrast to the majority of existing tree-based algorithms which determine split points when data instances bear equal weight, XGBoost can accommodate weighted data with its distributed weighted quantile sketch algorithm [92].

Block Structure for Parallel Learning: To expedite computational processes, XGBoost can leverage multiple CPU cores through its block structure system design, which permits data layout reuse in subsequent iterations, thus facilitating split finding and column sub-sampling.

Cache Awareness: To cope with non-continuous memory access necessitated for the retrieval of gradient statistics, XGBoost employs internal buffers in each thread for the storage of these statistics.

Out-of-Core Computing: This feature optimizes the use of available disk space for the effective management of voluminous datasets that exceed memory capacity.

XGBoost confers numerous advantages:

- Performance: XGBoost consistently yields high-caliber outcomes across a diverse range of machine learning tasks, rendering it a preferred choice among winners of Kaggle competitions.
- Scalability: The design of XGBoost emphasizes on efficacious and scalable training, rendering it suitable for handling large datasets.
- Customizability: XGBoost provides a broad array of tunable hyperparameters, thus allowing a heightened level of customization.
- Handling of Missing Values: XGBoost innately manages missing values, which streamlines the process of working with real-world data.
- Interpretability: XGBoost proffers data pertaining to feature importance, facilitating a better understanding of the variables that significantly influence prediction-making.

However, XGBoost also possesses certain disadvantages:

- Computational Complexity: The algorithm can be resource-demanding, particularly with larger models, which may be less ideal for systems with constrained resources.
- Potential for Overfitting: XGBoost may overfit when trained on smaller datasets or when an excessive number of trees are utilized.
- Hyperparameter Tuning: While XGBoost's hyperparameters offer substantial customization, tuning them necessitates expertise and can be time-intensive.
- Memory Requirements: Working with large datasets can be memory-intensive, rendering XGBoost potentially unsuitable for systems with limited memory capacity.

4.5.2. XGBoost Hyperparameters

XGBoost, a robust machine learning algorithm, employs an array of parameters categorized into general parameters, booster parameters, and learning task parameters, to optimize and regulate the behavior of its model. These parameters are indispensable in managing the model's complexity, mitigating overfitting, and enabling expedited convergence, particularly in scenarios exhibiting high class imbalance [93].

The "learning_rate" parameter bears similarity to the learning rate in Gradient Boosting Machines (GBM). This parameter induces a shrinkage effect on the weights at each successive step to avert overfitting. From a technical standpoint, after each boosting step, we multiply the weights of the features by this factor. The study under consideration employs a learning rate of 0.1.

The "max_depth" parameter regulates the maximum depth of any tree within the model, akin to GBM. It represents a crucial parameter to preclude overfitting as elevated depth values might cause the model to discern relations that are exceedingly specific to individual instances. The maximum depth in the study is set at 7.

"gamma", also denoted as "minimum loss reduction", is a regularization parameter. It stipulates the minimum loss reduction required to effect an additional partition on a leaf node of the tree. A larger gamma imparts a more conservative nature to the algorithm. In thestudy, gamma is assigned a value of 1.

"subsample" defines the fraction of the total observations to be randomly sampled for each tree. Lower values of subsample can render the model more conservative, thereby precluding overfitting. However, extremely low values could potentially lead to underfitting. In the study, subsample is assigned a value of 0.8, indicating that 80% of the data instances are utilized for constructing each tree.

"colsample_bytree" and "colsample_bylevel" dictate the fraction of columns to be randomly sampled for each tree and for each split at every level, respectively. In the study, "colsample_bytree" is set to 0.5, implying that half of the columns are sampled at each tree.

"reg_lambda" signifies the L2 regularization term on weights, a mechanism employed to counteract overfitting. It is set to 10 in the study.

"scale_pos_weight" is leveraged in circumstances characterized by high class imbalance as it aids in faster convergence. In the study, "scale_pos_weight" is set to 1, suggesting no specific class imbalance.

In combination, all these parameters synergistically enhance the performance of the XGBoost model, providing a balance between bias and variance, and ensuring that the model generalizes effectively to unseen data.

4.5.3. Pseudocode

Procedure XGBoost_Training(data, labels, num_rounds, learning_rate, max_depth):

```
    Initialize model with a constant prediction value
    For i = 1 to num_rounds:
        residuals = Compute_Residuals(data, labels, model)
        tree = Build_Tree(data, residuals, max_depth)
        Update model with tree weighted by learning_rate
    EndFor
    Return model

Procedure Compute_Residuals(data, labels, model):
    predicted_values = Predict(model, data)
    residuals = labels - predicted_values
    Return residuals

Procedure Build_Tree(data, residuals, max_depth):
    If max_depth is reached or other stopping criteria met:
        Return leaf_node, which is the mean of residuals
    Else:
        best_split = Find_Best_Split(data, residuals)
        left_tree = Build_Tree(data left of best_split, residuals left of best_split,
max_depth-1)
        right_tree = Build_Tree(data right of best_split, residuals right of best_split,
max_depth-1)
        Return node(best_split, left_tree, right_tree)

Procedure Find_Best_Split(data, residuals):
```

```
    For each feature in data:
        Consider potential split points, compute gain (using residuals)
    EndFor
    Return split that maximizes gain

Procedure Predict(model, sample):
    value = constant_prediction_value
    For each tree in model:
        value += learning_rate * Tree_Prediction(tree, sample)
    EndFor
    Return value

Procedure Tree_Prediction(tree, sample):
    If tree is a leaf:
        Return the value of the leaf
    If sample meets criteria at node:
        Return Tree_Prediction(left_child_tree, sample)
    Else:
        Return Tree_Prediction(right_child_tree, sample)
```

5. MODELS

5.1. Non Connected Model

Within the realm of study, two distinct paradigms is discerned: the non-connected model and its connected counterpart.

In the non-connected model framework, the emphasis is predominantly placed on the contemporaneous state, implying that only the present data point is evaluated, devoid of any reference to preceding or subsequent states. This paradigm operates on a foundational premise wherein datasets, even though collected in a temporal sequence (24 discrete datasets representing each hour for the respective cohorts), remain autonomous entities, uninfluenced by their temporal neighbors. Thus, these datasets are bereft of any mutual dependencies or interconnections. This approach aligns with traditional methodologies in data analysis, which advocate for an independent assessment of each dataset within the sequence. Hence, each hourly dataset undergoes a singular, isolated analysis.

To quantify the efficacy and robustness of various computational algorithms within this modeling framework, an assortment of algorithms—including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Multi-Layer Perceptrons (MLP), Random Forests (RF), and eXtreme Gradient Boosting (XGBoost)—were rigorously evaluated in relation to their performance metrics across the designated cohorts. Elucidated in Figure 3, the study scrutinized three distinct patient cohorts. These cohorts were characterized by 24 individual datasets, each one encapsulating hourly data spanning the inaugural day subsequent to the patient's admission into the Intensive Care Unit (ICU). When accounting for the comprehensive assessment, a total of 360 experimental outcomes were procured. This expansive number was the resultant product of applying the aforementioned quintet of algorithms across the all of datasets.

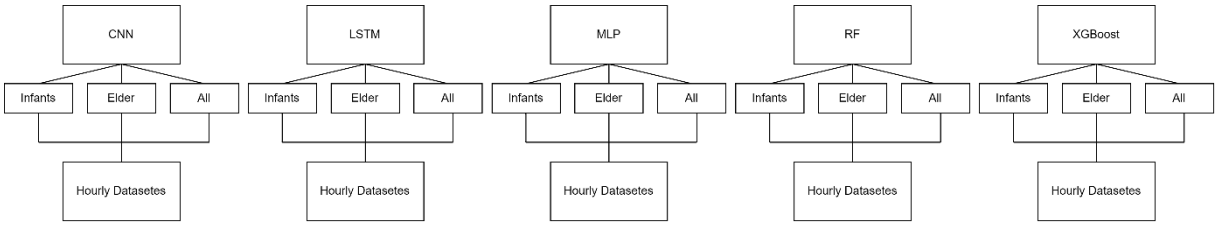

Figure 23. Methodology of Non-Connected Model

5.2. Connected Model

The idea of "*Information from a previous hour could be useful in the early alert of sepsis by considering actual hour*" was inspired by the development of the connected model. The conceptualization that data derived from a preceding temporal period might be instrumental in enhancing the early detection mechanisms for sepsis, especially when assessed in juxtaposition with the present hour, finds its origins in the evolution of the interconnected model. This model posits the significance of historical data as a potential catalyst in refining predictive capabilities, particularly in critical health conditions such as sepsis. The underpinning rationale is rooted in the recognition that temporal antecedents, when systematically incorporated, can augment the model's precision, ensuring timely interventions.

The development of the connected model has been instrumental in underpinning the hypothesis that data derived from a preceding temporal point could augment the early detection mechanisms for sepsis when assessed in the context of the current time point. The connected model is defined by its capacity to incorporate the prior probability of an instance belonging to a specific class, colloquially termed the "confidence level," at a given time point, denoted as t , as an innovative feature for the successive dataset for each instance at the subsequent temporal point, $t+1$.

Such an approach is predicated on the idea of infusing insights pertaining to the antecedent state of the system into the predictive framework for its current state. The

inherent strength of this methodology lies in its potential to exploit the temporal interdependencies that manifest within a dataset. This attribute is of paramount importance in scenarios involving time series data. A quintessential exemplar of its applicability is in the realm of healthcare, wherein the overarching objective is to prognosticate the onset of pathologies like sepsis by critically evaluating a patient's vital sign trajectory over a span of time.

Intrinsically, the likelihood of a patient being categorized under a specific cohort carries a concomitant degree of confidence. By integrating the confidence level elicited from antecedent data as an avant-garde feature for imminent datasets in a chronologically sequenced manner, the model is endowed with an enriched informational base. This, in turn, empowers the model to enhance the precision of its predictions, whilst simultaneously having the capability to identify and rectify anomalies present within the training dataset. Such a model is of particular significance in scenarios where the probability associated with class affiliations is projected to exhibit temporal fluctuations. In these cases, the assimilation of this probability as a distinct feature potentially augments the model's predictive accuracy [10].

$$t_0 < att_1, att_2, att_3 \ldots att_n > p_0$$
$$t_1 < att_1, att_2, att_3 \ldots att_n, p_0 > p_1$$
$$t_n < att_1, att_2, att_3 \ldots att_n, p_{n-1} > p_n$$

The notation delineated above represents a structured sequence of tuples. Each tuple, represented as t_0, encapsulates attributes from att_1 through att_n. Successive to this, the tuple t_n encompasses identical attributes ranging from att_1 to att_n , further augmented with an additional attribute p_{n-1}. For instance, the tuple t_1 incorporates the same attributes att_1 to att_n, and is supplemented with an additional attribute p_0, which is indicative of the confidence level pertaining to t_0.

It is imperative to recognize that the mathematical methodologies employed to derive the confidence levels across various algorithms might exhibit disparities. When provided a

sequential array of input features represented as $att = [att_1, att_2, att_3 \ldots att_n]$, wherein *n* signifies the sequence's cardinality, the resultant outputs of the Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) architectures can be collectively symbolized by a vector h_n. This vector pertinently denotes the hidden state of the terminal timestep. The computation for determining the confidence level is thus represented as:

$$p = sigmoid\ (w^n h_n + b)$$

In this equation, w stands as the weight vector associated with the terminal neuron present in the MLP, LSTM, and CNN architectures. The symbol *b* represents a bias term introduced to the equation. The activation function employed, named "sigmoid", and exhibits a characteristic sigmoidal curve, ensuring the output is confined between 0 and 1, rendering it apt for representing a confidence level.

In the domain of ensemble learning, when presented with an input feature vector, denoted as *att* , each individual decision tree within the Random Forest (RF) ensemble renders a classification prediction, which is either categorized as positive or negative. To ascertain the confidence level associated with the prediction, one can compute the proportion of decision trees within the ensemble that align with the positive class designation. This relationship can be mathematically articulated as:

$$p = \frac{n_{pos}}{n_{trees}}$$

In the above equation, the variable n_{pos} embodies the aggregate of decision trees within the RF ensemble that discerns the input feature vector att as aligning with the positive class. Concurrently, n_{trees} demarcates the total number of decision trees encapsulated within the ensemble.

Venturing into the domain of gradient boosting, given an input feature vector *att*, the eXtreme Gradient Boosting (XGBoost) algorithm meticulously crafts a collection of decision trees. These trees, in tandem, calculate a confidence score. This score undergoes a subsequent transformation, morphing it into a probabilistic value by invoking the logistic function. This transformation can be encapsulated in the following mathematical representation:

$$score = \sum(w_i \times f_i(x)$$

$$p = \frac{1}{1 + e^{-score}}$$

In the aforementioned formulation, $f_i(x)$ epitomizes the prediction elicited from the i-th decision tree within the ensemble, while w_i symbolizes the weightage ascribed to the said i-th tree. Additionally, the base of the natural logarithm, *e*, is invoked within the exponential function. The resultant confidence level, p , emerges as the output from the logistic transformation, effectively transmuting the score into a probability that resides within the interval [0,1].

Figure 24 provides a visual representation that illuminates the feed-forward mechanism of hourly datasets, wherein the confidence level is innovatively integrated as a distinct feature. This portrayal is particularly salient within the confines of the study, accentuating the essence of the interconnected model in enhancing prediction capabilities.

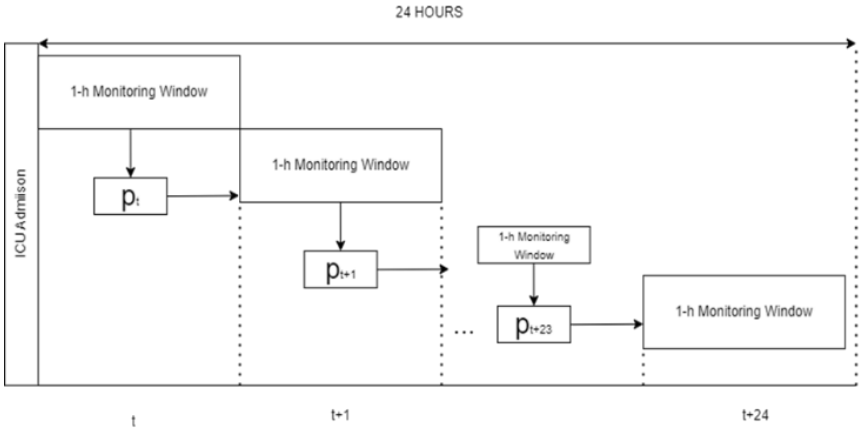

Figure 24. Connected Model – The Feed Forwarding of Hourly Data Set with Confidence Level

Algorithm and psuedo code fpr proposed model for the research [10]

Step #1: Divide the dataset at time t into 5 folds.

folds = create_folds(dataset_t, 5)

Step #2: For each fold, do the following:

for fold in folds:

a. Split the fold into training and test sets.
   train_data, test_data = split_fold(fold)

b. Train a classifier (XGBoost, RF, MLP, LSTM, CNN) on the training set, using the actual features of the dataset.
   classifier = train_classifier(train_data)

c. Use the trained classifier to predict the class labels for the test set instances.
   test_predictions = predict_labels(classifier, test_data)

d. For each instance in the test set, calculate the confidence level, as predicted by the classifier.
   probs = predict_probabilities(classifier, instance)

Step #3 . Use the confidence level as a new feature in the dataset at time t+1.

dataset_tplus1 = add_feature(dataset_tplus1, "confidence_level", confidence_level)

Step #4. Train a classifier on the dataset at time t+1, including the new feature, confidence level, from time t.

classifier_tplus1 = train_classifier(dataset_tplus1)

Step# 5. Use the trained classifier to classify instances in the dataset at time t+1.

tplus1_predictions = predict_labels(classifier_tplus1, dataset_tplus1)

6. RESULTS

6.1. Infant Cohort

In the investigation into the infant cohort, an array of machine learning models was evaluated based on their predictive accuracy measured by F1 scores. The cohort consisted of 2243 infants identified with sepsis (minority class) and 5857 sepsis-negative infants (majority class). This disproportion between the classes, typical in such datasets, may affect models' performance, potentially skewing results towards the majority class.

Table 20. F1 Results for Infant Cohort

F1 Score		Maximum	Average	Minimum
Connected Model	CNN	0.841	0.724	0.507
	LSTM	0.529	0.073	0.000
	MLP	0.653	0.583	0.484
	RF	0.899	0.882	0.855
	XGBoost	0.889	0.877	0.862
Non-Connected Model	CNN	0.627	0.569	0.460
	LSTM	0.447	0.049	0.000
	MLP	0.725	0.619	0.513
	RF	0.894	0.881	0.858
	XGBoost	0.859	0.889	0.877

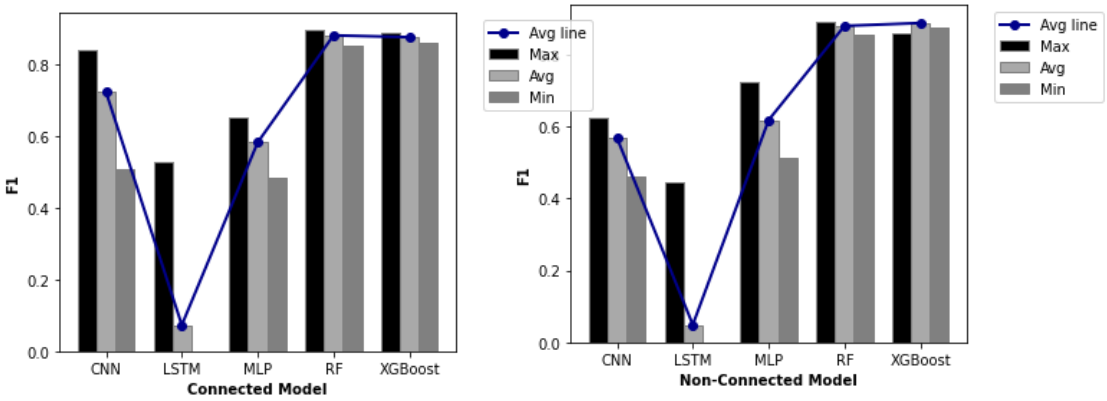

Figure 25. F1 Results of Infant Cohot

The analysis revealed that the connected Convolutional Neural Networks (CNN) model routinely secured F1 scores above 0.5, with many surpassing 0.7, highlighting its consistent predictive reliability. In contrast, the connected Long Short-Term Memory (LSTM) model, despite its peak F1 score of 0.529, held an average of only 0.073, failing to meet expected performance benchmarks. The connected Multi-Layer Perceptron (MLP) model's performance was acceptable, with scores generally hovering around 0.6, while the connected versions of Random Forest (RF) and Extreme Gradient Boosting (XGBoost) consistently achieved scores above 0.8, indicating their superior prediction precision and sensitivity.

A similar performance trend was noted in non-connected models. The CNN and MLP models secured scores mostly above 0.5. However, the non-connected LSTM model's mean F1 score was even lower at 0.049. Both the non-connected RF and XGBoost models showcased exemplary performance, with scores usually surpassing 0.8.

A marked enhancement in the average F1 score for the connected CNN model over its non-connected counterpart was evident in Table 20. Given ample data and parameter tuning, CNN models are renowned for their advanced learning and generalization accuracy. This performance can be further amplified by incorporating probability and confidence measures. It was also observed that both the RF and XGBoost models

maintained comparable average F1 scores in both connected and non-connected configurations.

Interestingly, the connected MLP model performed below its non-connected counterpart, suggesting that the inherent class disparity negatively impacted MLPs. During the training phase, a noticeable performance surge was seen for the predominant class, while the minority class's performance diminished.

Although the LSTM model is intricate in design, it struggled to handle data imbalance, emphasizing its challenges in learning and generalizing effectively from a limited and uneven dataset. This challenge can be ascribed to the limited number of sepsis-positive cases. Predicting sepsis accurately remains a daunting task, but specific machine learning models hint at potential advancements in this area.

Table 21. Metrics for Infant Cohort

Connected Model						**Non-Connected Model**				
	CNN	**LSTM**	**MLP**	**RF**	**XGBoost**	**CNN**	**LSTM**	**MLP**	**RF**	**XGBoost**
Hours	t4	t5	t4	t5	t5	t14	t3	t23	t8	t7
Precision	0.800	0.554	0.716	0.879	0.876	0.748	0.635	0.655	0.873	0.887
Recall	0.795	0.507	0.557	0.891	0.891	0.539	0.345	0.810	0.902	0.891
F1	0.797	0.530	0.627	0.885	0.883	0.627	0.447	0.725	0.887	0.889
Accuracy	0.889	0.745	0.812	0.935	0.933	0.819	0.759	0.826	0.935	0.937
AUC	0.858	0.673	0.735	0.921	0.920	0.733	0.633	0.821	0.925	0.923
Specificity	0.922	0.839	0.913	0.951	0.950	0.929	0.922	0.832	0.948	0.955

In the examination of the infant cohort, various machine learning models were meticulously assessed for their abilities to detect sepsis. The analysis encompassed both

connected and non-connected versions of five models: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Multi-Layer Perceptron (MLP), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost).

6.1.1. Connected Models:

1. CNN: Sepsis identification was accomplished at the 4th hour, exhibiting a precision of 0.800, sensitivity of 0.795, an F1 score of 0.797, an overall accuracy rate of 0.889, an Area Under the Curve (AUC) of 0.858, and a specificity of 0.922. The model thus demonstrated commendable efficacy in early detection and balanced performance.

2. LSTM: Detected sepsis at the 5th hour with a precision of 0.554, sensitivity of 0.507, an F1 score of 0.530, an overall accuracy of 0.745, and an AUC of 0.673, reflecting a somewhat limited performance.

3. MLP: Identified sepsis at the 5th hour, with a precision of 0.716, sensitivity of 0.557, an F1 score of 0.627, overall accuracy of 0.812, AUC of 0.735, and specificity of 0.913, indicating an average classification ability.

4. RF: Detected sepsis at the 5th hour, with strong precision (0.879), sensitivity (0.891), F1 score (0.885), overall accuracy (0.935), AUC (0.921), and specificity (0.951), indicating an exemplary performance.

5. XGBoost: Identified sepsis at the 5th hour, aligning closely with RF's metrics, further highlighting the model's proficiency in sepsis identification.

6.1.2. Non-Connected Models:

1. LSTM: Detected sepsis at the 3rd hour, but with restricted sensitivity (0.345), F1 score (0.447), and otherwise moderate specificity (0.922) and overall accuracy (0.759).

2. MLP: Identified sepsis at the 23rd hour, with sensitivity (81.0%), precision (65.5%), F1 score (0.725), specificity (83.2%), overall accuracy (82.6%), and an AUC value (0.821), verifying its accurate diagnostic abilities.

3. RF: Detected sepsis at the 8th hour, with high metrics: sensitivity (90.2%), precision (87.3%), F1 score (0.887), AUC (0.925), specificity (94.8%), and overall accuracy (93.5%).

4. XGBoost: Detected sepsis at the 7th hour, showing similar high values as the non-connected RF model.

The findings underscore that connected models generally demonstrate enhanced effectiveness in early sepsis detection (t=4 and 5 hours) compared to non-connected models. However, this comparison might be considered biased in the case of the LSTM model due to the low F1 score of the non-connected version. The latter group tends to detect sepsis later, between the 3rd and 23rd hours, although the performance metrics of some non-connected models (such as MLP) are poor.

The implications of this research are profound. The connected model algorithms' ability to leverage early hour data and detect sepsis earlier can potentially improve therapeutic outcomes and hold life-saving consequences. The data suggests that the development and refinement of these models may lead to increased precision and reduced false positives while maintaining sensitivity. Further research is warranted in this direction to continue to enhance the early diagnostic capabilities of these models in the context of sepsis.

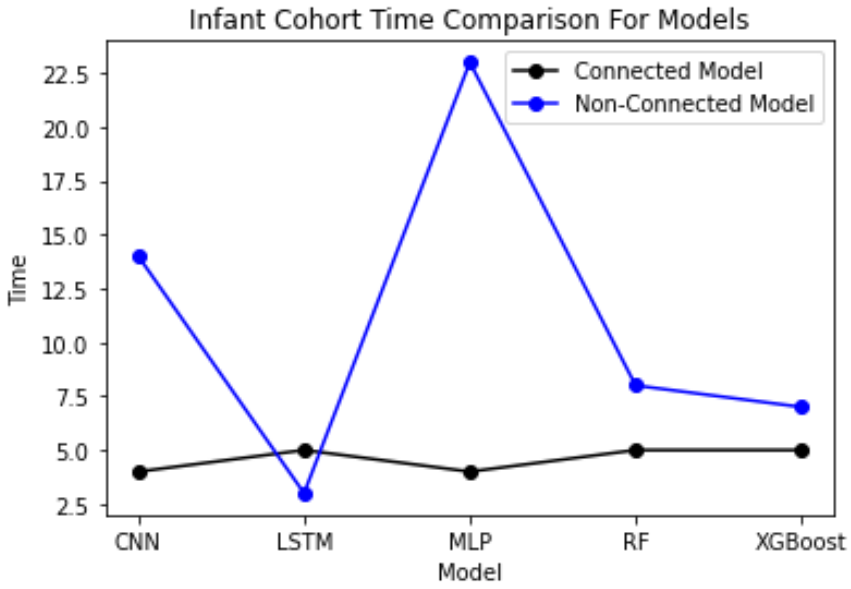

Figure 26. Time Comparison for Infant Cohort

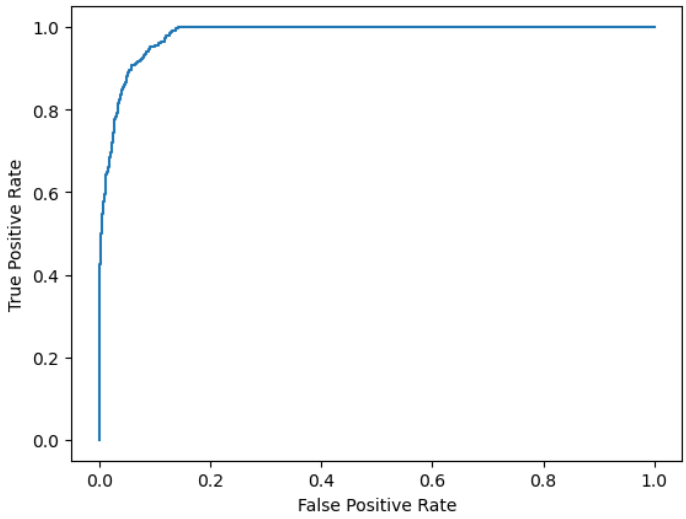

Figure 27. XGBoost ROC for Infant Cohort

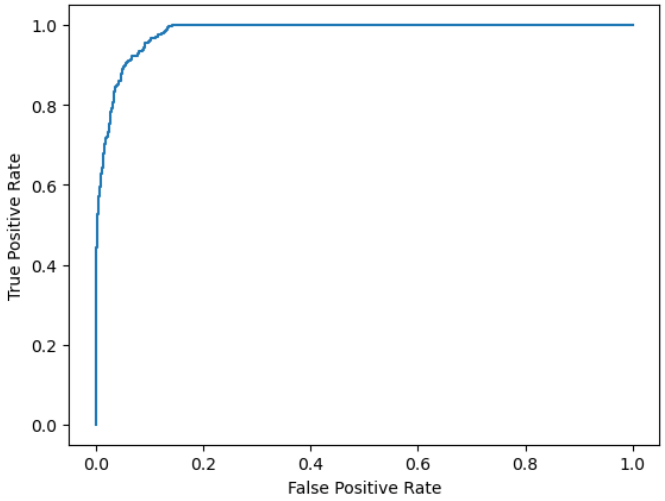

Figure 28 RF ROC for Infant Cohort

XGBoost Connected MAE: 0.066

Random Forest Connected MAE: 0.069

6.2. Elder Cohort

In the study focused on the elderly cohort, a diverse performance spectrum was observed among the evaluated models. The connected Convolutional Neural Network (CNN) model exhibited an F1 score ranging from 0.779 to 0.877, averaging at 0.831, highlighting its dependable capacity to yield sensitive and precise predictions. In contrast, the connected Long Short-Term Memory (LSTM) model presented an F1 score band between 0.771 and 0.784, suggesting consistent yet marginally subdued performance. The connected Multilayer Perceptron (MLP) model delineated an average F1 score of 0.783, fluctuating between 0.749 and 0.811, showcasing its conventional prowess in harmonizing sensitivity and precision. Remarkably, the connected Random Forest (RF) and Extreme Gradient Boosting (XGBoost) models both exhibited stellar performance with the RF model having an almost unwavering F1 score of 0.935 and the XGBoost model recording scores oscillating between 0.933 and 0.935.

Table 22. F1 Results for Elder Cohort

F1 Score		**Maximum**	**Average**	**Minimum**
Connected Model	**CNN**	0.877	0.831	0.779

	LSTM	0.784	0.771	0.753
	MLP	0.811	0.783	0.749
	RF	0.935	0.935	0.934
	XGBoost	0.935	0.933	0.93
Non-Connected Model	**CNN**	0.879	0.848	0.791
	LSTM	0.783	0.772	0.743
	MLP	0.814	0.777	0.748
	RF	0.999	0.996	0.992
	XGBoost	0.932	0.935	0.934

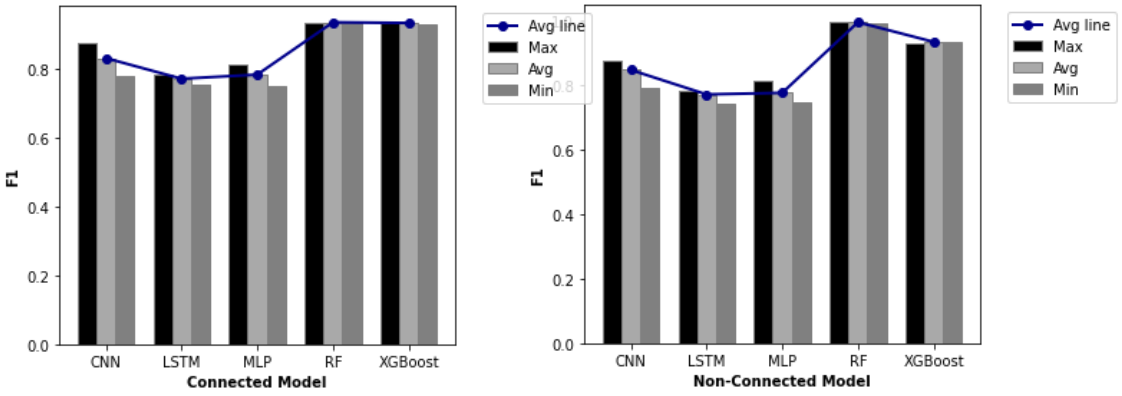

Figure 29. F1 Results of Elder Cohort

In scrutinizing the non-connected model variants within this cohort, the non-connected CNN model yielded an F1 score interval of 0.791 to 0.879, with an average value of 0.848. Despite its propensity to uphold a commendable balance of sensitivity and precision, some variance in performance was noted. The non-connected LSTM model posted an average F1 score of 0.772, spanning 0.743 to 0.783, which, while respectable, lagged slightly behind other models. The non-connected MLP model registered an average F1 score of 0.777, veering between 0.748 and 0.814, emblematic of its adeptness at achieving equilibrium between sensitivity and precision. Notably, the non-connected RF model eclipsed other non-connected counterparts, with an impressively consistent F1

score straddling 0.992 to 0.999. The non-connected XGBoost model also held its ground, tabulating an average F1 score fluctuating between 0.932 and 0.934.

A comparative appraisal revealed that F1 scores of the connected and non-connected CNN models were closely aligned. The connected LSTM variant marginally surpassed its non-connected counterpart in maximum and average F1 scores, notwithstanding comparable minimum scores. This contrast was more pronounced for the MLP models, where the connected iteration significantly overshadowed the non-connected one.

When diagnosing sepsis in the elderly cohort, models leveraging ensemble learning strategies, particularly the connected RF and XGBoost models, manifested exceptionally high F1 scores across the board, underscoring their adeptness in intricate classification challenges. In the realm of non-connected models, the RF variant took the lead with the highest average and peak F1 scores. While the majority of other models registered comparatively modest scores, the non-connected CNN and XGBoost models, with their relatively elevated peak F1 scores, showcased potential under particular circumstances.

Interpreting the outcomes, it is discernible that a model's efficacy is influenced by the foundational machine learning principles it adopts and the specific configurations and parameters set. For instance, CNNs, owing to their proclivity for discerning patterns and features, register commendable performance metrics in both connected and non-connected scenarios. LSTM models, being specialized in identifying temporal patterns, exhibit reliable performance, albeit slightly lagging behind CNNs. This distinction might be attributed to sensitivities around parameter configurations and data preprocessing intricacies inherent to LSTMs.

In a similar vein, MLPs, while recognized for pattern identification capabilities, occasionally grapple with complex data, which might elucidate their relatively diminished F1 scores. In stark contrast, the stellar metrics of both RF and XGBoost models can likely be credited to their ensemble-based design. The ensemble technique's

ability to discern intricate patterns and manage varied data types and scales appears beneficial, especially for diagnosing sepsis in the elderly demographic.

Table 23. Metrics for Elder Cohort

Connected Model						**Non-Connected Model**				
	CNN	**LSTM**	**MLP**	**RF**	**XGBoost**	**CNN**	**LSTM**	**MLP**	**RF**	**XGBoost**
Hour	t3	t16	t6	t3	t3	t10	t12	t4	t7	t6
Precision	0.808	0.671	0.778	0.878	0.873	0.884	0.669	0.725	0.88	0.883
Recall	0.886	0.91	0.846	0.999	0.995	0.868	0.916	0.84	0.997	0.994
F1	0.845	0.771	0.811	0.935	0.93	0.876	0.772	0.777	0.935	0.935
Accuracy	0.79	0.654	0.746	0.91	0.905	0.842	0.653	0.69	0.911	0.911
AUC	0.752	0.551	0.705	0.874	0.871	0.831	0.548	0.63	0.876	0.878
Specificity	0.619	0.193	0.564	0.75	0.746	0.795	0.18	0.42	0.755	0.761

The investigation into the performance of various models in identifying sepsis within distinct time intervals yielded significant insights. In the context of the elderly cohort:

6.2.1. Connected Models

1. CNN: Detection was achieved at t=3 hours, illustrating a sensitivity of 80.8%, confirming the accuracy of 80.8% of its positive predictions. Additionally, this model displayed a sensitivity rate of 88.6%, predicting 88.6% of positive cases. The F1 score was 0.845, indicating a satisfactory balance between sensitivity and precision. However, an AUC value of 0.752 suggests a slightly increased likelihood of false-positive predictions. Its specificity was recorded at 61.9%.

2. LSTM: This model displayed potential for detection at t=16 hours. Sensitivity and specificity values were 0.671 and 0.193, respectively. This translates to accurate

prediction of 67.1% of positive cases, but the model showed increased chances of predicting false positives.

3. MLP: Detection took place at t=6 hours, with precision and sensitivity rates standing at 84.6% and 77.8%. The F1 score reached 0.811, with a specificity of 56.4%.

4. RF: Exhibiting proficiency, this model detected sepsis at t=3 hours with an F1 score of 0.935, a sensitivity of 87.8%, and precision at 99.9%. It also showed a robust AUC of 0.874 and a specificity of 75%.

5. XGBoost: Detection was observed at t=3 hours. Metrics recorded were a sensitivity of 99.5%, precision of 87.3%, F1 score of 0.93, specificity of 74.6%, and an AUC of 0.871.

6.2.2. Non-Connected Model

1. CNN : This model identified sepsis at t=10 hours, revealing a sensitivity of 86.8%, precision of 88.4%, F1 score of 0.876, AUC of 0.831, and specificity of 79.5%.

2. LSTM : Detectable at t=12 hours, this model boasted a sensitivity of 91.6% and precision of 66.9%, but its specificity stood low at 18%.

3. MLP: Operating at t=4 hours, the metrics were a sensitivity of 84%, precision of 72.5%, F1 score of 0.777, AUC of 0.63, and specificity of 42%.

4. XGBoost - RF: Detection was achieved at t=6 hours, with a sensitivity of 99.4%, precision of 88.3%, F1 score of 0.935, specificity of 76.1%, and an AUC of 0.878.

Among these, the most proficient models in overall performance were the connected and non-connected RF and XGBoost. Notably, LSTM-based models encountered challenges with false positives due to low specificity, which could have serious implications in

scenarios like patient care where false positives in conditions such as sepsis have significant repercussions.

Furthermore, connected models, particularly CNN, RF, and XGBoost, consistently detected sepsis symptoms earlier (at t=3 hours) than their non-connected counterparts. This points to the potential advantage of connected models in facilitating swift sepsis detection within the elder cohort.

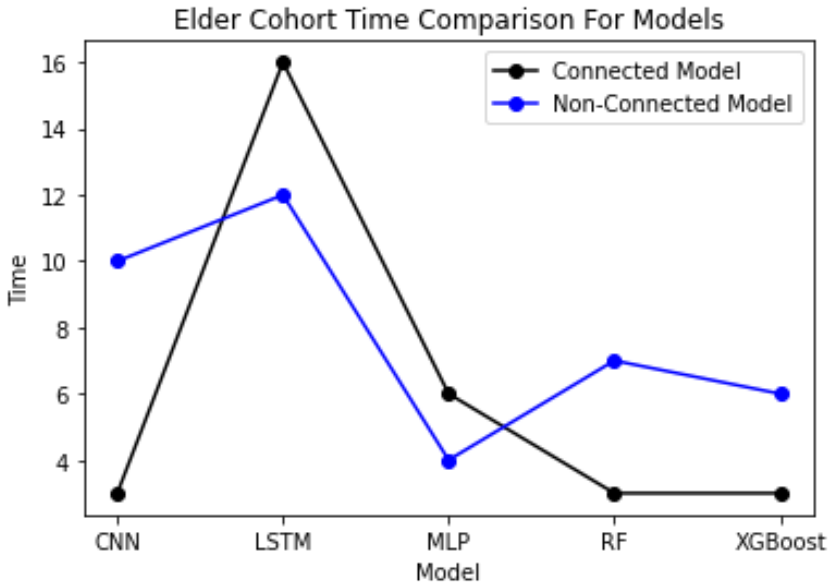

Figure 30. Time Comparison for Elder Cohort

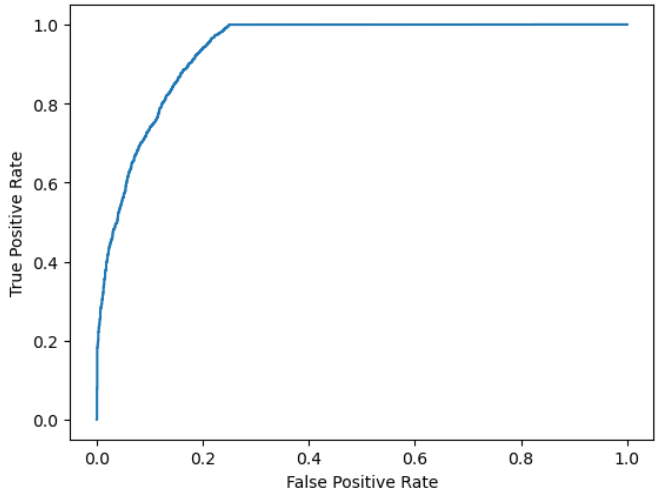

Figure 31. RF ROC for Elder Cohort

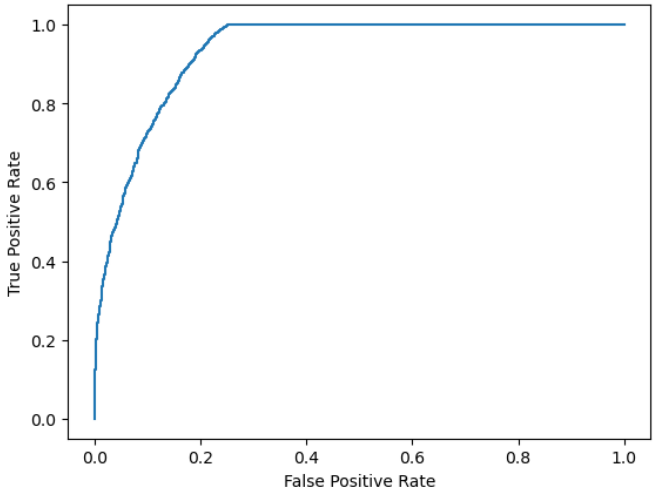

Figure 32. XGBoost ROC for Elder Cohort

XGBoost Connected MAE: 0.0907

Random Forest Connected MAE: 0.0898

6.3. All Age Cohort

In an examination of the all-age cohort for the detection of sepsis, a comparative analysis of various machine learning models was conducted, elucidating their respective performances. Through this analysis, a range of F1 scores emerged, reflecting the capabilities and characteristics of each model.

Table 24. F1 Results for All Age Cohort

F1 Score		Maximum	Average	Minimum
Connected Model	**CNN**	0.899	0.841	0.749
	LSTM	0.752	0.658	0.72
	MLP	0.794	0.751	0.684
	RF	0.916	0.914	0.911
	XGBoost	0.917	0.914	0.911
Non-Connected Model	**CNN**	0.853	0.83	0.749
	LSTM	0.753	0.733	0.697
	MLP	0.784	0.749	0.688
	RF	0.916	0.914	0.911
	XGBoost	0.911	0.917	0.914

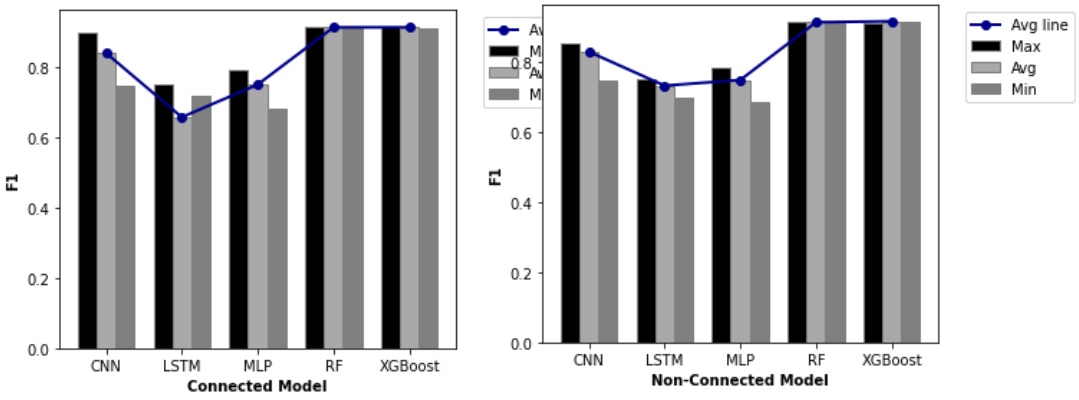

Figure 33. F1 Results of All Age Cohort

The connected Convolutional Neural Network (CNN) model revealed a commendable balance, achieving a maximum F1 score of 0.899, a minimum score of 0.749, and an average score of 0.841. This highlighted its proficiency in accurately identifying true positives. Conversely, the connected Long Short-Term Memory (LSTM) model, with a maximum F1 score of 0.752, a minimum score of 0.658, and an average score of 0.719, underperformed in comparison to the connected CNN model. The connected MLP model showed a reasonable equilibrium, reaching peak F1 ratings of 0.793, dipping down to 0.684 at the lowest, and averaging at 0.751. When it comes to connected algorithms, both the RF and XGBoost models outperformed others, attaining mean F1 scores of 0.914 and 0.917, respectively. Both algorithms hit a top score of over 0.91 and never dipped below this mark.

In the realm of non-connected models, different performances were observed. While the non-connected CNN model exhibited an average score of 0.830, the non-connected LSTM model's average was slightly lower at 0.733. The non-connected MLP model achieved an average F1 score of 0.749, whereas the non-connected RF and XGBoost models outperformed others, with scores of 0.916 and 0.917, respectively.

An evaluation of these models based on their F1 scores established the connected RF, connected XGBoost, non-connected RF, and non-connected XGBoost models as high performers. Their success in predicting positives and identifying true positives was highlighted by the close approximation of their maximum, minimum, and average F1 scores. Conversely, both connected and non-connected CNN and MLP models, though performing admirably on average, trailed the RF and XGBoost models, resulting in somewhat inferior performance. Connected and non-connected LSTM models were noted to generally underperform when compared to others.

The comparison between connected and non-connected versions of the same models unveiled remarkably similar performances. Instances were identified where non-connected versions slightly outperformed their connected counterparts, such as the non-connected LSTM model. However, RF models, both connected and non-connected,

manifested remarkably similar performances, revealing no apparent advantage of one over the other.

The collective results affirm that the connected RF, XGB, non-connected RF, and non-connected XGBoost models deliver the best performance in predicting positives and identifying true positives within the all-age cohort. Despite the commendable results of other models, they were found to exhibit a slightly inferior performance in comparison.

Several underlying intricacies and suitability factors in machine learning models for sepsis prediction in an all-age cohort were discerned through these results. The ensemble nature of RF and XGBoost algorithms, with their amalgamation of predictions from simpler models, contributed to their exceptional performance. In contrast, CNN models' somewhat inferior F1 scores were attributed to the problem's nature and the architecture's inherent strengths. The MLP models' sensitivity to parameter tuning and overfitting potential were considered influential in their performance, while the LSTMs' general underperformance was linked to their inherent complexity and training challenges. The robustness of machine learning models across different architectural settings was evident in the observed similarity between connected and non-connected models. Instances of non-connected models outperforming their connected counterparts were also noted.

The ensemble techniques (RF and XGBoost) showcased the most favorable performance. The nature of the task, complexity of the models, parameter settings, and overfitting tendencies were perceived as pivotal factors accounting for the observed performance variation. The selection of an appropriate model corresponding to the task's unique characteristics and requirements emerged as an essential consideration for achieving optimal results.

Based on the evaluation of F1 scores, the models demonstrating superior performance include the connected RF, connected XGB, non-connected RF, and non-connected XGB. The similarity in maximum, minimum, and average F1 scores for these models is not only evident but also commendably high, underscoring their aptitude in predicting and correctly identifying true positives.

While the connected and non-connected CNN models manifest commendable average F1 scores, their maximum and minimum scores tend to be inferior to those of the RF and XGB models. Generally, CNN models, irrespective of their connectivity status, exhibit a balanced performance but marginally trail behind the performance metrics of the connected RF and XGB models.

The MLP models, both connected and non-connected variants, showcase laudable average F1 scores. Nevertheless, their peak F1 scores are surpassed by those of the RF, CNN, and XGB models. Relative to the top-tier models, MLP models exhibit a minor shortfall in their capability to predict positives and accurately pinpoint true positives.

In assessing the performance of LSTM models, both connected and non-connected, a consistent trend of subpar performance emerges, especially when gauging average F1 scores. The recorded scores for LSTM models tend to be inferior compared to other model variants. Contrarily, the assertion that LSTM models often excel at predicting positives and recognizing true positives seems incongruous with the aforementioned observations.

A comparative analysis of models existing in both connected and non-connected versions reveals a marginal performance disparity between the two configurations. However, there are instances where the non-connected iteration demonstrates a slight edge, as observed in the superior performance of the non-connected LSTM model over its connected counterpart.

The F1 scores for the RF models, regardless of their connection status, are strikingly consistent. A meticulous examination of the average, maximum, and minimum F1 scores fails to spotlight any discernible differentiation between these two versions. Their performance metrics are almost mirror images, and neither can be conclusively favored over the other.

The foremost performers in predicting and accurately identifying true positives are the connected RF, XGB, and their non-connected counterparts. Conversely, while other models garner respectable outcomes overall, they tend to fall short of the standards set by the aforementioned models.

Upon evaluation of various machine learning models applied to sepsis prediction within an all-age cohort, distinct differences in performance were evident.

Table 25. Metrics for Elder Cohort

Connected Model						**Non-Connected Model**				
	CNN	**LSTM**	**MLP**	**RF**	**XGBoost**	**CNN**	**LSTM**	**MLP**	**RF**	**XGBoost**
Hour	t3	t9	t6	t3	t3	t9	t21	t10	t8	t8
Precision	0.885	0.716	0.772	0.844	0.848	0.843	0.688	0.783	0.854	0.859
Recall	0.874	0.745	0.812	0.991	0.985	0.851	0.831	0.777	0.983	0.974
F1	0.879	0.730	0.792	0.912	0.912	0.847	0.753	0.78	0.914	0.913
Accuracy	0.867	0.696	0.764	0.894	0.895	0.83	0.698	0.758	0.898	0.898
AUC	0.867	0.690	0.758	0.883	0.884	0.828	0.683	0.756	0.888	0.889
Specificity	0.860	0.636	0.705	0.775	0.783	0.804	0.534	0.734	0.793	0.803

6.3.1. Connected Models

The connected Convolutional Neural Network (CNN) model predicted sepsis as early as t=3 hours, manifesting a precision of 0.885, a sensitivity of 0.874, and an F1 score of 0.879. Additionally, a specificity score of 0.860 indicates a notable overall performance, indicative of a low false positive rate.

In contrast, the connected Long Short-Term Memory (LSTM) model forecasted sepsis at t=9 hours, achieving an F1 score of 0.730, a precision of 0.716, and a sensitivity of 0.745. Despite its capacity to predict true positives, it exhibited lower precision and sensitivity in comparison to the connected CNN model. This model's specificity value of 0.636 implies a higher number of false positives.

The connected Multilayer Perceptron (MLP) model predicted sepsis, reflecting an F1 score of 0.792, a precision of 0.772, and a sensitivity of 0.812, with a specificity of 0.705.

In ensemble methods, both the connected Random Forest (RF) and XGBoost models displayed superior performance, predicting sepsis at t=3 hours. Specifically, the RF model yielded an F1 score of 0.912, a precision of 0.844, and a remarkable sensitivity of 0.991, with a specificity of 0.775. The XGBoost model displayed analogous results.

6.3.2. Non-Connected Models

Regarding non-connected models, the non-connected CNN model anticipated sepsis at t=9 hours, with metrics including an F1 score of 0.847, a precision of 0.843, a sensitivity of 0.851, and a specificity of 0.804. Conversely, the non-connected LSTM model, at t=21 hours, achieved an F1 score of 0.753, a precision of 0.688, a sensitivity of 0.831, and a specificity of 0.534. Meanwhile, the non-connected MLP model predicted sepsis at t=10 hours, presenting an F1 score of 0.780, a precision of 0.783, a sensitivity of 0.777, and a specificity of 0.734.

The non-connected RF and XGBoost models, forecasting sepsis at t=8 hours, exhibited high sensitivities of 0.983 and 0.974, F1 scores of 0.914 and 0.913, and specificity values of 0.793 and 0.803, respectively.

The performance metrics indicate that the connected XGBoost and RF models boast the most superior F1 scores at 0.912. Furthermore, these models also manifest commendably elevated values in precision, sensitivity, and specificity. When evaluating F1 scores, the

non-connected iterations of RF and XGBoost models emerge as top contenders, registering scores of 0.914 and 0.913, respectively.

While the CNN models, in both connected and disconnected configurations, present robust F1 scores (0.879 and 0.847, respectively), they marginally lag behind the RF and XGBoost models. The F1 scores attributed to LSTM and MLP models are notably subdued in comparison. A subsequent appraisal of sensitivity and precision values for LSTM and MLP models accentuates this observation when juxtaposed against the CNN, RF, and XGB models. The connected LSTM model's metrics, with a precision of 0.716, sensitivity of 0.745, and an F1 score of 0.730, are slightly eclipsed by the non-connected LSTM model, which garners a precision of 0.688, sensitivity of 0.831, and an F1 score of 0.753.

The performance analysis between connected and non-connected MLP models unveiled marginal disparities, positioning them in a comparable performance bracket.

In the context of sepsis symptom prediction, RF and XGBoost models, encompassing both connected and non-connected variants, distinguished themselves as the most adept. A distinct advantage of the connected models is their ability to detect sepsis symptoms considerably earlier than their non-connected counterparts. Specifically, at the third hour post onset, connected CNN, RF, and XGBoost models displayed prowess in early symptom identification—a pivotal capability for the prompt assessment and initiation of requisite medical interventions. Contrarily, the non-connected LSTM model delineated a considerably delayed detection, only at the 21st hour. However, among the non-connected configurations, the RF and XGB models exhibited an earlier detection at the 8th hour.

When evaluating based on F1 scores, the connected XGBoost and RF models achieved the highest scores of 0.912, with non-connected RF and XGBoost models following closely at 0.914 and 0.913, respectively. Despite their commendable performance, the

CNN models, both variants, scored slightly lower, while the LSTM and MLP models recorded lower F1 scores and decreased precision and sensitivity.

To encapsulate, this examination suggests that connected models generally excel in early identification of patients manifesting sepsis symptoms. Their combined attributes of early detection complemented by high precision, sensitivity, and specificity render them particularly conducive for the swift and accurate recognition of sepsis indicators.

Performance consistency between connected and non-connected models suggests their inherent efficacy for sepsis prediction. However, connected models tend to anticipate sepsis symptoms earlier. Among connected models, the CNN, RF, and XGBoost discerned sepsis at t=3 hours, whereas among non-connected counterparts, the RF and XGBoost detected it earliest at t=8 hours. This data highlights the increased precision, sensitivity, and specificity of connected models, underscoring their potential in facilitating early sepsis detection, which can lead to timely therapeutic interventions.

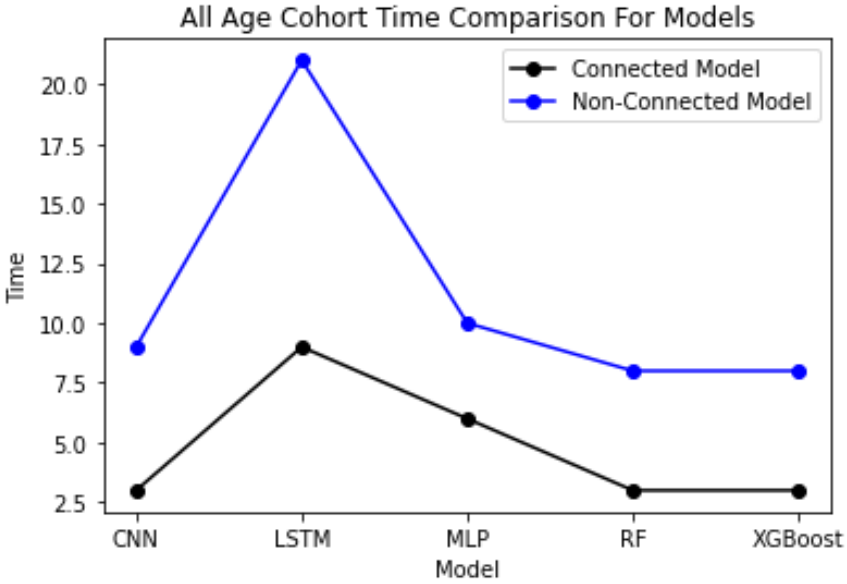

Figure 34. Time Comparison for All Age Cohort

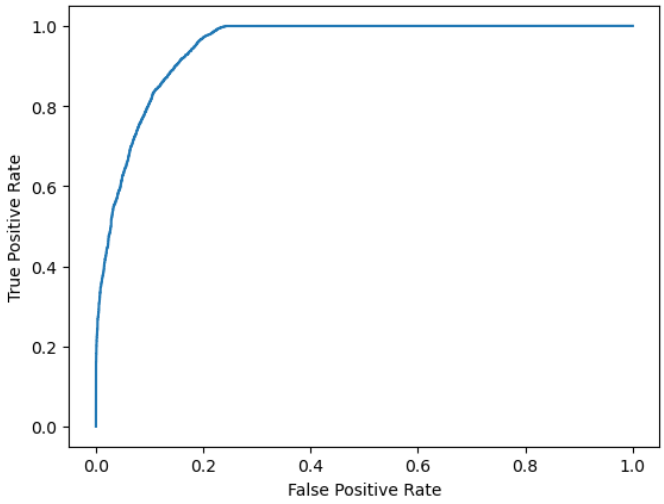

Figure 35. XGBoost ROC for All Age Cohort

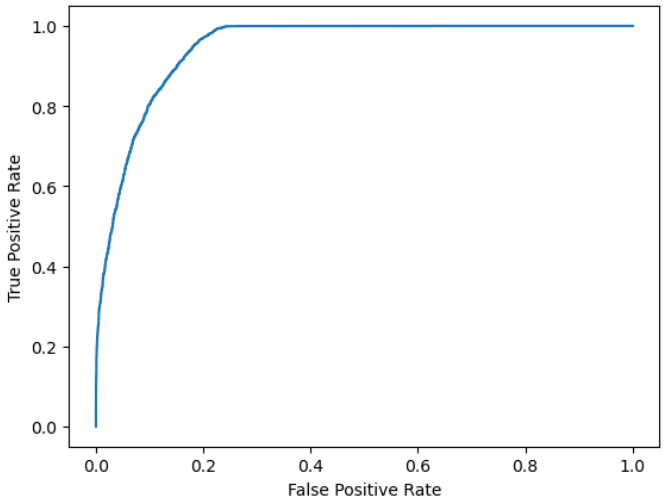

Figure 36. RF ROC for All Age Cohort

XGBoost Connected MAE: 0.10644348744616884

Random Forest Connected MAE: 0.10538717802876411

7. DISCUSSION

Sepsis, a swiftly progressing and often fatal complication stemming from infections, has been the crux of numerous studies due to the imperative need for timely diagnosis and

intervention. The intrinsic dynamics of this condition make early detection a non-negotiable component of effective medical intervention. Given this context, the present study took an innovative approach, leveraging the vast swathes of data in the MIMIC-III database to underline the role of "time" and connected machine learning models in the diagnostic journey.

One cannot emphasize enough the significance of "time" in the realm of sepsis management. Recognizing the acute trajectory of sepsis, this research went beyond conventional methods by fragmenting the MIMIC-III dataset into hourly segments during the initial 24-hour span of a patient's ICU stay. By doing so, the study shone light on the nuanced changes occurring in patient conditions – changes that are often the precursors to a potentially severe septic episode.

The segmentations across all ages, elder, and infant cohorts provided unique insights. Each age group showcased distinct challenges and patterns related to sepsis onset. The variations in metrics such as heart rate, temperature, and others across cohorts signaled the intricate and multifaceted nature of sepsis presentation and progression, underscoring the necessity of age-specific diagnostic tools and interventions.

The study's standout finding was the superior performance of connected models. Rooted in a holistic approach, connected models, by their very design, ensure that the progression of diseases like sepsis is not treated as isolated, disconnected events. The implication here is vast – by understanding and leveraging past data, current diagnostic predictions become more refined and accurate. Such an approach is especially salient for swift-progressing ailments, where the past can offer vital clues about the present and future trajectory of the disease.

However, while the connected models clearly outshone their non-connected counterparts, the value of models like RF and XGBoost in certain situations cannot be understated.

These simpler models might still serve as useful tools, especially in situations demanding rapid predictions or when there's a paucity of extensive patient history.

The interlinking of time, patient demographics, and machine learning techniques presents a formidable front in the ongoing battle against sepsis. This study, through its innovative methodologies, reiterates the criticality of early detection and lays down a path for future explorations. With continuous advancements, the hope is to develop an optimized, accurate, and holistic predictive model, bringing transformative changes to the way sepsis is diagnosed and managed in intensive care settings.

In the progression of this study, there was a meticulous exploration and application of diverse machine learning techniques. The core objective of these methods was to amplify the ability to provide early alerts for sepsis, focusing especially on the nuances between non-connected and connected data models.

The study implemented two prominent decision tree-based algorithms - Random Forest (RF) and XGBoost. While RF utilized a forest of 500 decision trees to diversify predictions, XGBoost was designed on the gradient boosting framework. In the latter, subsequent trees aim to mitigate the errors identified in the preceding ones [25,26].

The prowess of artificial neural networks was tapped into through Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN). MLP was the base model, whereas LSTM, with its unique gating mechanism, targeted time series data. CNN, complemented by convolutional layers, was employed for hierarchical feature identification. Optimal model performance was ensured by judiciously choosing initialization methods, activation functions, and dropout techniques [27-29].

The study shed light on the stark differences between the two models. The non-connected model examined data points as separate entities, not tapping into the temporal patterns. The connected model, however, added a dimension by integrating the confidence level

from previous data to the current prediction, highlighting the richness of temporal data, especially in predicting conditions like sepsis.

The methods were backed by sound mathematical reasoning. Functions like sigmoid activation, logistic functions, and tree count ratio were systematically employed across the models to transform raw data into relevant confidence scores.

Yet, the research acknowledges the intrinsic limitations. RF and XGBoost might grapple with extensively high-dimensional data. In contrast, deep learning methods, albeit powerful, can demand substantial computational resources and data for training. The connected model, though innovative, can occasionally be biased due to its reliance on previous data instances.

This research sets the stage for forthcoming endeavors. Future scholars might want to merge the strengths of traditional and deep learning models or explore more intricate temporal modeling methods. It's clear that as computational tools evolve and data becomes increasingly accessible, the onus is on researchers to mold these resources into tools for enhanced patient care.

7.1. Infant Cohort

In the context of sepsis prediction within the infant cohort, this study elucidates crucial insights about the performance of various connected and non-connected machine learning models, emphasizing the diverse ways in which these models respond to the imbalanced data and distinct characteristics of the task.

The connected models, particularly CNN, RF, and XGBoost, have displayed significant promise for early sepsis detection, a vital component for timely diagnosis and intervention. The connected CNN model's proficiency in early sepsis detection, identifying cases as early as the 4th hour, is particularly noteworthy, as is the consistent

performance of the connected RF and XGBoost models, which maintained F1 scores above 0.8. However, the underperformance of the connected LSTM and MLP models highlights challenges related to data imbalance and model complexity, emphasizing the need for careful consideration in model selection and tuning.

The non-connected models, particularly the RF and XGBoost, demonstrated consistent accuracy and reliability, reinforcing their robustness. Conversely, the non-connected LSTM's lower mean F1 score indicates suboptimal performance, despite its ability to detect sepsis at the 3rd hour. These variations in performance suggest that non-connected models may have unique applications in specific clinical scenarios.

A direct comparison between connected and non-connected models reveals that connections allow for enhanced pattern recognition and data synthesis in some cases, such as the CNN algorithm. This superior learning capability is reflected in the connected CNN model's improvement over its non-connected counterpart. The nuance in the results for other models like MLP, where the non-connected model fared slightly better, suggests the need for further exploration.

The imbalanced nature of the dataset has likely played a pivotal role in model performance, creating biases towards the majority class. The study underscores the necessity for future research to focus on techniques to counterbalance this imbalance, such as oversampling the minority class or utilizing weighted loss functions.

Different models exhibited varied efficacies in diagnosis time, an essential factor in clinical settings. The early diagnosis capability of some models might still render them valuable in specific scenarios where timely detection is prioritized over other performance metrics.

The study highlights that reliance on a single performance metric can be misleading. Balancing precision, sensitivity, and other metrics like the F1 score is integral to ensure

a model's applicability in a clinical context, where false negatives can have serious ramifications.

The findings recommend further research to explore ensemble techniques that combine the strengths of both connected and non-connected models, alongside strategies to handle imbalanced datasets. Such advancements could contribute to the development of more reliable tools for early sepsis detection.

In conclusion, this comprehensive analysis of machine learning models in predicting sepsis within an infant cohort not only underscores the potential of machine learning in healthcare but also emphasizes the intricate nature of model selection and implementation. The nuanced findings offer essential guidance for practitioners, suggesting both the promise and challenges inherent in applying these technologies for early disease detection. Further refinement and research into these models may pave the way for improved precision, reduced false positives, and maintained sensitivity, ultimately contributing to enhanced patient care within the delicate and complex domain of infant sepsis diagnosis.

7.2. Elder Cohort

In the quest to optimize sepsis diagnosis for the elderly population, it is imperative to evaluate the efficacy of different machine learning models. The elder cohort, with its unique challenges, demands tailored solutions that ensure both accuracy and timeliness.

Among the connected models tested, the CNN model demonstrated commendable results, with consistent high F1 scores indicative of its precise and sensitive predictive capabilities. In contrast, the LSTM model registered a narrower spectrum of F1 scores, suggesting some constraints in its performance range. The MLP model, however, adeptly balanced sensitivity and precision, thereby cementing its position as a reliable choice. Yet, it was the RF and XGBoost models that truly distinguished themselves, consistently clocking high F1 scores and heralding their superior acumen in sepsis detection. A

parallel trend was evident among the non-connected models: the CNN and LSTM models demonstrated aligned and fair performance, respectively, while the MLP and RF models exhibited consistent strength, with the XGBoost model joining their ranks in showcasing robust performance. Interestingly, a head-to-head comparison revealed negligible differences in F1 scores between connected and non-connected variants of the CNN models, with the connected LSTM and MLP models edging out their non-connected counterparts.

Diving deeper into sepsis detection within the elder cohort, the RF and XGBoost models, both in their connected and non-connected configurations, showcased exemplary F1 scores. Their balanced performance across various metrics, namely sensitivity, precision, and specificity, underscores their potential as frontline tools in elder sepsis detection. However, the connected models held a distinct advantage in early detection, with the CNN, RF, and XGBoost variants identifying sepsis symptoms as early as the 3rd hour. This rapid detection capability can be a game-changer in the management of time-sensitive conditions like sepsis. Yet, a note of caution is warranted. LSTM models, despite their potential, grappled with false positives, evidenced by their low specificity. Such limitations have dire implications in clinical practice, risking unnecessary treatments and potential patient distress.

It's evident that the connected RF, XGBoost, and to some extent, the CNN models, hold considerable promise for the early detection of sepsis in the elderly. These findings accentuate the importance of strategic model selection, with an emphasis on those that maximize both rapid detection and diagnostic accuracy. Future endeavors might delve deeper into refining these models, potentially exploring hybrid solutions that amalgamate the strengths of individual models. For instance, the integration of the CNN's adeptness in feature extraction with the RF's prowess in classification might lead to enhanced diagnostic tools. Furthermore, addressing the limitations of LSTM models, especially their propensity for false positives, should be a priority.

This study's findings illuminate the performance intricacies of various connected and non-connected machine learning models in sepsis diagnosis for the elderly. While significant advancements have been made, there's an undeniable need for continual research and model refinement. The overarching objective remains steadfast: harnessing AI's power to ensure efficient, timely, and precise patient care for our elder cohort, a population whose well-being demands both our respect and our best technological innovations.

7.3. All Age Cohort

The integration of machine learning techniques into medical diagnosis holds transformative potential, offering healthcare practitioners tools to enhance patient outcomes. This is especially pivotal for conditions like sepsis, where prompt and accurate diagnosis is paramount. In analyzing the performances across various models for the all age cohort, several key insights emerge:

Foremost, the RF and XGBoost models, in their connected configurations, have emerged as frontrunners. Their consistent high F1 scores underline their capacity to balance both precision and recall adeptly, essential metrics in the domain of medical diagnostics. Given the critical nature of diagnosing sepsis, these models' robustness accentuates their promise as invaluable diagnostic aids in clinical settings. This juxtaposes against the relatively underwhelming performance of LSTM models, suggesting that while they excel in capturing long-term dependencies in sequential data, they might not be optimally aligned with this dataset's nature or the specific challenges of sepsis detection.

Across the board, connected models demonstrated an edge over their non-connected counterparts. Their superior early prediction abilities and enhanced metric scores reinforce the hypothesis that integrating information between models can lead to a more comprehensive data representation and, in turn, improved prediction capabilities.

The criticality of early detection in sepsis cannot be understated. Highlighting this, connected models, especially the CNN, RF, and XGBoost variants, demonstrated the

prowess to foresee sepsis symptoms as early as the third hour. Such capability holds immense clinical significance, given that timely interventions in sepsis cases often delineate the thin line between patient recovery and deterioration.

The dual challenges in sepsis detection lie in identifying genuine cases (sensitivity) and avoiding false alarms (specificity). While high sensitivity values ensure patients with sepsis are not overlooked, robust specificity reduces the risk of misdiagnosis, which can lead to unwarranted treatments. The connected LSTM model's reduced specificity underscores the need for caution, pointing towards potential areas of model refinement.

Beyond architectural distinctions, the data's characteristics, training methodologies, and inherent noise and outliers play pivotal roles in dictating model efficacy. The resilience of ensemble models like RF and XGBoost against noise, outliers, and class imbalances speaks to their adaptability, making them well-suited for real-world clinical data riddled with intricacies.

The evident performance disparities emphasize the nuanced choice researchers and clinicians face. It's not merely about gravitating towards high F1 scores; it's about comprehensively weighing sensitivity against specificity, early detection against absolute accuracy, and model interpretability against performance metrics. Added to this is the necessity to consider model integration with existing healthcare systems, external validation on diverse datasets, and ethical considerations intrinsic to AI-driven diagnostics.

The results unveiled in this study herald an exciting juncture in the integration of machine learning with sepsis diagnostics for an all age cohort. Ensemble models, notably the RF and XGBoost, along with the CNN, stand out, showcasing their adeptness at harnessing intricate data patterns for precise, timely predictions. However, as we tread forward, the choice of models, training methodologies, and external validations will dictate the real-world impact of these tools. The journey ahead beckons a refined interplay of data science

and clinical expertise, aimed at optimizing patient outcomes in the face of challenges like sepsis.

The investigation offers a pivotal contribution to the understanding of how connected machine learning models can be instrumental in the early detection of sepsis, emphasizing its applicability across diverse age cohorts, notably infants, the elderly, and a general age demographic.

1. Emphasis on Temporal Dynamics: One of the most salient findings is the augmented prediction capability afforded by accounting for temporal dependencies. This emphasizes the quintessential nature of time-series data in medical diagnosis, as it can often illuminate predictive insights that traditional static data models might overlook.

2. Comparative Efficacy of Models: A granular exploration of the machine learning models revealed nuances in their performances.

- CNN: Demonstrating its prowess, the connected CNN variant consistently advanced diagnosis times across all cohorts, highlighting its potential in time-critical conditions like sepsis.

- LSTM: While it showcased strengths in processing sequential data for the general cohort, its efficacy dwindled for infants and the elderly. Such divergence underscores the possibility of age-specific data intricacies influencing model outcomes.

- MLP: The connected MLP model displayed differential efficacy across the cohorts, hinting at the model's potential challenges in discerning complex temporal patterns inherent to certain age groups.

- RF & XGBoost: These models stood out in their connected configurations, with commendable AUC scores accentuating their capacity for distinguishing sepsis instances with precision.

3. Model Suitability for Different Age Cohorts: The research underscores that a one-size-fits-all approach may not be optimal. Certain models, such as the connected CNN for

infants, may be more aligned with specific age cohort dynamics, thus warranting tailored model choices based on patient demographics.

4. Research Scope and Limitations: The study's anchoring to the MIMIC III dataset poses potential limitations. Diverse datasets, reflecting a spectrum of clinical realities, may present nuanced patterns and challenges. Therefore, the replicability and validity of our findings necessitate broader data evaluations.

5. Clinical Ramifications: The models' capacity to advance sepsis diagnosis by several hours is not just statistically significant but also holds profound clinical value. In the critical landscape of sepsis management, these augmented prediction horizons can substantially modulate patient outcomes.

6. Cautionary Notes on Model Interpretation: Despite the innovative foray of introducing connected models for sepsis prediction, prudence remains imperative. Regular validations, coupled with methods like cross-validation, can mitigate risks of overfitting and ensure robust model performances.

This study heralds a promising avenue in the use of connected machine learning models for sepsis prediction, potentially marking a paradigm shift in clinical interventions. The transformative potential of these models, especially in advancing prediction times, is undeniably profound. However, as we chart this promising trajectory, a comprehensive exploration and validation across diversified datasets remain essential to fortify the findings and ensure their broad-scale applicability in real-world clinical environments.

8. GENERAL EVALUATION AND KEY DISCUSSION POINTS

In ICUs, sepsis and its complications contribute to approximately 42% of mortalities [1]. Globally, sepsis affects an estimated 30 million individuals each year, both directly and indirectly. Those who recover from sepsis face risks of long-term health complications,

such as chronic morbidities and permanent disabilities. Given these alarming statistics and observations, there exists a critical need to explore how AI technologies can be optimized to establish early diagnosis of sepsis and determine effective treatment strategies. A delayed diagnosis of sepsis substantially increases the risk of organ dysfunction and death. At this juncture, AI-based early warning systems emerge as a potential solution. Leveraging advanced data analytics and algorithms, these systems can promptly alert clinicians when there are adverse changes in patient parameters. Early warning systems not only have the potential to preserve patients' lives but can also reduce hospital mortality rates and enhance patients' quality of life. Consequently, the role of AI in sepsis diagnosis and treatment is foundational.

The core research question focuses on how suspicions of sepsis can be detected more accurately and earlier. Acknowledging the paramount importance of early diagnosis for sepsis, attention has been given to the initial 24-hour period of patients in ICUs. This phase is critical from both diagnostic and therapeutic perspectives, and decisions made within this timeframe can profoundly influence a patient's prognosis. In a typical clinical setting, by the end of 24th hour of a patient's stay in the ICU, the required laboratory and physiological data for diagnosis have usually been collected. At this stage, pharmacotherapy has been initiated, and its efficacy has undergone rigorous evaluation by clinicians. As a result, the period following the initial 24 hours encompasses the patient's diagnostic and therapeutic processes under thorough clinical oversight. Optimizing pharmacotherapeutic treatment protocols for sepsis patients necessitates a complex decision-making process, based on the severity and presence of sepsis. The study aims to provide a guiding framework for clinicians during this pivotal decision-making process, especially within the initial 24 hours. In this regard, patients' hourly clinical assessments have been approached using two distinct modeling strategies. In the "Connected Model" approach, a patient's clinical data at a specific hour is evaluated cumulatively with the data from preceding hours. Conversely, in the "Non-Connected Model" approach, only the clinical indicators of that specific hour are taken into consideration. These two modeling techniques are methodologies routinely implemented by clinicians. The research aims to integrate these methods with AI algorithms to offer a more systematized evaluation method.

Early warning systems are central to the research in question. Such systems aim to rapidly identify suspicions of sepsis, particularly within the initial 24 hours, thereby granting clinicians and opportunity for proactive intervention. Yet, every research endeavor carries inherent methodological limitations. Machine learning and AI-based sepsis studies face several significant challenges, notably the scope of the study and the heterogeneity of data sources. In particular, many studies possess limited sample sizes, weakening the generalizability of outcomes. Furthermore, sourcing datasets from a single institution and relying on retrospective study designs can constrain the generalization capabilities of algorithms. Sepsis research typically advances along two main axes: contributions from datasets and methodological innovation. In this investigation, the MIMIC-III dataset was chosen with an emphasis on enhancing the methodological approach, though this selection also introduced specific constraints. Primarily, the MIMIC-III database does not provide direct hourly patient data. While it encompasses detailed information about ICU patients, measurement and test outcomes are presented in a "time-stamped" manner. This suggests additional steps are required to access hourly data. During this process, two tables named "sepsis3" and "sepsis3_cohort" were created. While "sepsis3" contains general information and diagnoses, "sepsis3_cohort" was designed for sepsis symptoms and indicators. Subsequently, hourly data for ICU patients was gathered using the "sepsis.all_icustays" table. This data was merged with sepsis suspicion labels. While MIMIC-III does not specify sepsis onset times, potential onset times for each individual were determined via SOFA scores.

To emphasise, the primary objective of this research was to establish a comprehensive hourly dataset for sepsis diagnosis and treatment. Consequently, an hourly dataset was developed integrating hourly features, sepsis labels, and other clinical indicators for use in sepsis diagnosis. It should be noted that while MIMIC-III offers time-stamped data, it does not provide hourly specific diagnoses. In this context, the study focused intensely on the most critical and uncertain initial 24 hours for every ICU patient.

The patient population was categorized into three age groups. Literature indicates that individuals under the age of 1 and those above 60 are more susceptible to sepsis.

Accordingly, all age, elderly, and infant groups were established. Datasets representing the 24-hour period for the first ICU day were prepared for these groups.

While MIMIC-III provides patients' SOFA scores, it does not specify specific sepsis onset times (“onset time”). To address this omission, the evolution of SOFA scores over time was examined in the hourly datasets, enabling the estimation of a possible sepsis onset time for each patient.

The dataset in use encompasses 16 critical variables. Among these, identifiers and time details such as “icustay_id”, “intime”, and “outtime” are comprehensively available. Moreover, the “sepsislabel” has been determined via the SOFA score, and age information (“age”) is consistently provided. Alongside basic information, the dataset incorporates vital signs, laboratory results, and specialized columns like “suspected_infection_time_poe”. Inevitably, dataset has missing values. While some tests occur twice daily, certain measurements are undertaken every minute. To address missing data, the most recent measurement served as the primary reference. However, in the absence of the latest measurement, the age group's mean was employed to populate the missing data.

Conducted descriptive statistical analyses reveal significant differences in fundamental statistical parameters, such as demographic diversity, variance, and standard deviation, across different age cohorts. Notably, analyses centered on the elderly cohort observed lesser variance and standard deviation values for certain biomarkers and clinical parameters. This observation might indicate that the elderly population possesses a more homogeneous demographic structure. Conversely, the "All Ages" category displays broader demographic diversity, resulting in higher standard deviation values for statistical parameters.

In the pediatric cohort, considering physiological development stages, more constrained variance and standard deviation values were procured for some parameters. However, data deficiencies were also identified for certain measurements in this cohort.

Chi-square analyses conducted for health parameters specific to different age cohorts determined the level of statistical significance as 0.01. The presence of statistical significance affirms that an observed effect or association is not random and verifies the existence of a meaningful relationship at a specific confidence level. With this analysis, the p-values obtained for the "Age" variable are notably low (<0.01), establishing that this variable displays statistically significant differences among distinct patient cohorts. Yet, the p-values derived for the "White Blood Cell Count", "ICU entry timestamp", and "ICU discharge timestamp" variables exceed 0.01, suggesting no statistically significant differences among different cohorts for these variables. Some vital parameters are statistically significant only for the "All Ages" cohort, whereas the “Creatinine” level has been found significant across all cohorts.

Although the SOFA score is not directly employed for sepsis diagnosis, it's pertinent to note that in clinical practice, a specific SOFA score threshold is recognized as an indication of sepsis. In this research, the objective was to validate the relationship between ICD-9 and the SOFA score using a diagnostic test table, employing the SOFA score for identification purposes. The high sensitivity and negative predictive value obtained signify that these two parameters aren't used interchangeably, yet they possess a strong correlation.

In situations governed by time-sensitive critical factors, the adopted connected modeling approach enriches feature sets, taking into account specific “confidence levels”. This method accommodates temporal connections between datasets, fostering more consistent and accurate predictions by the model. This research bases its methodology on two distinct modeling paradigms: Non-Connected and Connected. The non-connected modeling paradigm treats each data point as an isolated event. This implies that the model doesn't account for the preceding or subsequent hourly status of a patient, predominantly focusing on instantaneous states. In this perspective, each dataset is evaluated independently, devoid of temporal links. The non-connected modeling approach was implemented on three different patient groups. Each group comprises 24 distinct datasets,

which in total were analyzed using five different algorithms. This yielded 360 experimental datasets.

In the connnected modeling approach, questions arose about how effectively the disease trajectory, especially for critical conditions related to time such as sepsis, can be predicted. This method allows for a more holistic evaluation facilitated by the incorporated confidence levels.

To determine the statistical significance of the performance difference between the two models, the Wilcoxon test was employed. A statistically significant result from the Wilcoxon test might suggest superiority of one model over the other. A non-significant result from the test indicates comparable performance between the models. Ultimately, when setting the threshold value at 0.01, no statistically notable difference was discerned between the models. However, considering the time factor, the "connected" model delivered results in the 4th hour, while the "non-connected" model did so in the 14th hour.

This research demonstrates that using connected modeling paradigms effectively models the dynamic components of clinical processes. Among the algorithms, CNN, RF, and XGBoost exhibited superior performance compared to others. The specified t3, t4, and t5 time slots were determined as critical for algorithmic performance. In conclusion, with non-connected models, there's an opportunity to diagnose sepsis within the initial six hours, offering significant time savings for both patients and healthcare professionals. In the prevailing literature, a marked preference for non-connected models is evident. However, this study underscores the potential of connected modeling paradigms to make significant contributions to the field.

The research demonstrates that basing machine learning algorithms on a connected paradigm to model dynamic clinical processes can offer an effective alternative to current non-connected methodologies. The accuracy rates concerning early warning systems in the existing literature are congruent with the findings of this study.

A distinguishing feature of this research compared to similar studies in the literature is the utilization of the MIMIC-III dataset without any exclusion criteria. Employing datasets refined through exclusion criteria can potentially enhance model performance. Notably, while most studies in the literature employing the MIMIC III dataset tend to focus on 5,000- 10,000 patients, this research analyzes data from 60,000 patients, observing comparable performance. This suggests that the findings of this study are generalizable. Despite the creation of large datasets, the prevalent practice has been to downsize them primarily for model training. Such an approach was not adopted in this research, which might enhance the model's efficacy across a broader patient population. Foregoing exclusion criteria serves as an advantage in preventing unrealistic outcomes. Such an approach bolsters confidence in the generalizability of the research.

From a methodological perspective, the SOFA formula was chosen to determine the “onset time”. This approach, without resorting to complex statistical analyses, facilitates swift and effective outcomes. In academic literature, the performance of early warning systems often appears inferior compared to standard diagnostic methods. The findings of this research resonate with this trend.

In conclusion, it's posited that the connected model paradigm contributes significantly to the literature. This study advocates for connected modeling and contrasts it with the non-connected approach. The results indicate that the connected modeling approach presents advantages in diagnosing sepsis in early hours.

9. CONCLUSION

In the evolving landscape of medical research, the challenges posed by sepsis remain paramount, chiefly accentuated by its profound mortality implications, especially within the purview of intensive care units. As the medical community endeavors to address the formidable task of early sepsis detection, the need for timely diagnosis becomes even more poignant, given the drastic spike in mortality with each passing treatment hour. While the annals of research are replete with diverse methodologies, from canonical scoring systems like APACHE II, SIRS, and qSOFA to avant-garde machine learning

paradigms, it's evident that the quest for an optimal predictive tool is a dynamic and multifaceted journey.

Our current exploration stands at the nexus of sepsis research and state-of-the-art computational techniques. Venturing beyond traditional methods, we've tapped into the transformative potential of the "connected model" – an innovative approach that amalgamates machine learning with probabilistic constructs, seeking to harness historical patient data for robust predictions. The marked efficacy of this model across various algorithms, including the likes of MLP, LSTM, CNN, RF, and XGBoost, and its unparalleled ability to preempt sepsis onset in critical timeframes for different patient cohorts, showcases its immense clinical promise. Furthermore, by casting a spotlight on age cohorts with heightened sepsis vulnerability, namely the elderly and infants, the research champions a targeted approach while simultaneously ensuring comprehensive applicability across all age groups.

In essence, the study not only underscores the transformative potential of melding cutting-edge computational methodologies with pressing clinical exigencies but also establishes the "connected model" as a promising beacon in the continuum of sepsis prediction tools. As the global health fraternity confronts the multifarious challenges of sepsis, the confluence of clinical sagacity and computational acumen emerges as a vital alliance. This promising fusion augurs well for future endeavors, holding the potential to reshape patient care paradigms, and ensuring that interventions are both timely and targeted. The ensuing discourse will delve deeper into the findings, implications, and the prospective trajectory of this pivotal research endeavor.

Navigating the intricate terrains of critical care, especially within the confines of Intensive Care Units (ICUs), demands precision, timeliness, and an unwavering commitment to patient outcomes. Such imperatives become even more pronounced when contending with relentless adversaries like sepsis, whose rapid progression can dictate the trajectory of patient recovery. Within this context, the current study embarked on a quest to harness the depths of the MIMIC-III database, a rich reservoir of critical care data spanning over a decade at the Beth Israel Deaconess Medical Center. By anchoring our investigations

between 2001 and 2012, we endeavored to unveil intricate patterns of sepsis progression, harnessing the granularity of the data to shed light on the pivotal role of time in shaping patient trajectories in the ICU.

The ensuing narrative is a synthesis of our key revelations, from the nuanced insights gleaned from partitioning data into hourly segments, capturing the essence of the first 24 hours of a patient's ICU sojourn, to the innovative juxtaposition of connected versus non-connected diagnostic approaches. These findings, complemented by a meticulous exploration of variable significance across patient cohorts and the pioneering incorporation of machine learning paradigms, collectively chart a path that underscores the inextricable intertwining of data-driven methodologies and clinical imperatives in the ICU.

While our findings underscore the compelling prospects of such a detailed approach, it's equally imperative to juxtapose these revelations against the backdrop of inherent research limitations, offering a balanced and holistic view of the study's contributions. Thus, as we delve deeper into the study's conclusions, we do so with an appreciation of the immense potential held by the MIMIC-III database, and the broader ramifications of our findings on shaping the future trajectory of critical care diagnostics and interventions.

The vast realm of healthcare analytics remains a burgeoning frontier, characterized by both its potential to revolutionize patient outcomes and its intricate challenges. Within this paradigm, the timely detection of critical conditions like sepsis stands out as a quintessential challenge, particularly in the high-stakes environment of Intensive Care Units (ICUs). The present investigation pivots on this very intersection, meticulously exploring the confluence of multiple machine learning and deep learning paradigms with the objective of elevating sepsis detection through comprehensive hourly ICU admission datasets.

As we embark on the concluding reflections of this study, we are reminded of the multifaceted approaches undertaken—from the traditional non-connected model that provides an elemental understanding of each algorithm's prowess in isolation, to the more

nuanced connected model which underscores the profound impact of historical data in predicting imminent health states. Beyond just methodology, the study's analytical depth, manifested in mathematical formulations, throws light on the intricate mechanics of integrating prior probabilities in future predictions.

This synthesis, as we will delve deeper into, not only amplifies our understanding of sepsis detection mechanisms but also offers a beacon for future research endeavors, underlining the imperativeness of bridging temporal analytics with predictive health models. As we navigate through the culmination of this research, let us reflect upon its key takeaways and the broader implications they hold for the future of critical care analytics.

9.1. Infant Cohort

The analysis of the infant cohort, as presented, offers a comprehensive understanding of the efficacy of various connected and non-connected machine learning models in sepsis detection. Crucially, the study underscored the significance of employing the F1 score as a robust metric to assess model precision and recall concurrently, thereby providing a holistic overview of a model's diagnostic capability.

The connected models, particularly RF, XGBoost, and CNN, exhibit a notable proficiency in sepsis detection in the early stages (t=4 and t=5 hours). Their high sensitivity, precision, F1 score, and AUC values emphasize their potential as reliable diagnostic tools for sepsis. Meanwhile, the non-connected models showcased mixed results, with RF and XGBoost outperforming others, but with a slightly delayed diagnosis time compared to their connected counterparts.

The class distribution (Impact of Data Imbalance), which is skewed towards non-sepsis cases, influences the model performance. This disparity brings forward the challenge of handling imbalanced datasets in medical diagnoses, as it can compromise the model's capability to predict the minority class (sepsis-positive cases). This effect was particularly evident in the LSTM model's performance, which underscores the importance of data balance in achieving optimal outcomes.

An interesting observation is the variation in the time at which different models identified sepsis. Early detection is vital in sepsis management, thus emphasizing the significance of models that can diagnose sepsis in its nascent stages. The connected CNN, RF, and XGBoost models offer promising outcomes in this regard.

Between the connected and non-connected models, it's evident that connectivity offers an advantage in terms of sepsis detection, particularly for the CNN model. However, for algorithms like RF and XGBoost, the performance remains comparably high irrespective of connectivity.

The RF, XGBoost, and CNN models, both in their connected and non-connected formats, stand out as promising tools for sepsis diagnosis in the infant cohort. Their high performance metrics combined with their ability to diagnose sepsis at early stages make them potentially invaluable tools in clinical settings.

In conclusion, the study furnishes pivotal insights into the potential of machine learning models in the early diagnosis of sepsis in infants. While certain models shine in their diagnostic prowess, the study also accentuates the challenges posed by data imbalances. Future research could delve deeper into strategies for addressing dataset imbalances or explore ensemble methods that leverage the strengths of multiple models. Moreover, while the models show promise, translating them to real-world clinical settings necessitates further validation through larger cohorts and diverse datasets.

9.2. Elder Cohort

In this study, various connected and non-connected models were assessed in terms of their effectiveness in predicting sepsis for the elder cohort. The analytical metrics deployed included the F1 Score, precision, recall, accuracy, AUC, specificity, and time required for early diagnosis. This multi-faceted evaluation allowed for a comprehensive understanding of the models' proficiency in handling this critical classification task.

Our observations reveal distinct trends in performance. Foremost, Random Forest (RF) and XGBoost consistently emerged as the top-performing models across both connected and non-connected settings. Their superior F1 scores and balance between precision and recall underscore their aptitude in sepsis diagnosis, a domain where achieving equilibrium between false positives and true positives is paramount. This trend is corroborated by their high AUC values, indicating a commendable trade-off between sensitivity and specificity.

It's noteworthy that connected models generally diagnosed sepsis earlier than their non-connected counterparts. As sepsis is a time-sensitive ailment, the importance of expedited diagnosis cannot be overstated; every hour can be pivotal for patient outcomes. The faster detection times exhibited by the connected models, specifically the CNN, RF, and XGBoost, further accentuate their clinical value.

However, not all models showcased exemplary performance. The LSTM-based models, in particular, faced challenges with specificity. The low specificity rates indicate a propensity to register false positives. In medical contexts like sepsis prediction, where false alarms can have tangible repercussions, this limitation is especially concerning.

In summary, while many of the evaluated models showed promise, RF and XGBoost, especially in their connected configurations, stand out as particularly adept at early sepsis detection in the elder cohort. Their combination of high accuracy, balanced sensitivity-specificity trade-off, and quick detection makes them prime candidates for further clinical exploration and potential integration into healthcare systems. Yet, it is also imperative to consider model limitations and the unique characteristics of the elder cohort when interpreting and applying these results in real-world settings. Further studies might delve deeper into the mechanisms driving these performances and explore opportunities to refine the less performative models.

9.3. All Age Cohort

This study provides an insightful exploration into the performance of various models in early sepsis diagnosis across the all age cohort. Through comprehensive evaluation metrics, we elucidate the strengths and weaknesses of each model, offering valuable guidance for practitioners and researchers alike.

From our findings, it is evident that RF and XGBoost models, both in connected and non-connected configurations, outperform other models in terms of F1 scores. Their superior performance implies their efficacy in striking a balance between precision and recall. Moreover, the high specificity of these models further accentuates their capacity to minimize false positives, which is crucial for critical medical diagnoses like sepsis.

Another notable observation pertains to the connected models' ability to predict sepsis symptoms at earlier stages compared to their non-connected counterparts. Early detection is paramount in medical interventions, especially for conditions like sepsis where timely treatment can significantly impact patient outcomes. The CNN, RF, and XGBoost connected models particularly stand out with their capacity to identify symptoms at the t=3 hour mark, emphasizing their potential role in clinical settings for rapid patient assessment.

However, it's worth mentioning the underperformance of the LSTM models, both connected and non-connected, in the context of this study. Their relatively lower F1 scores, as well as precision and recall values, suggest a need for further optimization or consideration of alternative architectures for sepsis prediction.

In juxtaposition, the CNN models, especially the connected ones, show promise with commendable F1 scores, even though they don't match the excellence of the RF and XGBoost models. Their early prediction capabilities and overall good balance further spotlight their potential utility.

In summation, our analysis underscores the pivotal role of machine learning models in advancing medical diagnostics. The RF and XGBoost models, given their impressive performance, hold considerable promise for early and precise sepsis detection, with connected models generally showcasing a propensity for timely intervention. As the medical community continues to embrace technological advancements, studies like these will be instrumental in guiding the choice and optimization of models, ultimately striving for improved patient care and outcomes.

This research illuminated the potential of employing connected machine learning models, including CNN, LSTM, MLP, RF, and XGBoost, to enable early and accurate prediction of sepsis across various patient demographics. Distinct advantages of these connected models over traditional, non-connected ones were observed, notably in their ability to harness the power of temporal dependencies within the data. By factoring in prior system states, these models enhanced predictive accuracy—crucial in life-threatening conditions where timely interventions can substantially alter outcomes.

Noteworthy findings include the notable efficiency of the connected CNN, XGBoost, and RF models in early sepsis prediction across all cohorts. Especially in elderly and all-age groups, sepsis was predicted as early as t=3 hours post-ICU admission, and at t=4 hours for infants. This starkly contrasts with their non-connected counterparts which, in most scenarios, exhibited a delayed diagnosis. Specifically, the connected XGBoost and RF models displayed remarkable proficiency in distinguishing between sepsis and non-sepsis cases, as evidenced by their high AUC values.

Despite the promising results, it is essential to interpret them with caution. The study solely drew on the MIMIC III dataset—a retrospective, single-center compilation—which inherently limits the generalizability of the findings. Therefore, for a more comprehensive and universal understanding, investigations involving larger, more diverse datasets across multiple centers are imperative.

Additionally, while connected models advance sepsis prediction, they may also introduce challenges such as potential overfitting. Instances where the model attributes low probability to a particular class can be indicative of either mislabeling or model overreach. Thus, regular monitoring, iterative refinements, and validation are necessary to maintain and enhance the models' performance over time.

In sum, this research underscores the transformative potential of connected machine learning models in the realm of early sepsis diagnosis. While the XGBoost, CNN, and RF models have displayed particularly promising results, it is quintessential to augment these insights with additional studies across diverse datasets. As the medical fraternity grapples with the pressing challenge of sepsis, the integration of sophisticated connected models could pave the way for a paradigm shift in early detection and intervention.

10. REFERENCES

[1] A. Lever, I. Mackenzie, Sepsis: definition, epidemiology, and diagnosis, BMJ 335 (2007) 879. https://doi.org/10.1136/bmj.39346.495880.AE.

[2] M. Fink, H. Warren, Strategies to improve drug development for sepsis, Nat. Rev. Drug Discov. 13 (2014) 741–758. https://doi.org/10.1038/nrd4368.

[3] J. Su, Z. Tong, S. Wu, F. Zhou, Q. Chen, Research Progress of DcR3 in the Diagnosis and Treatment of Sepsis, Int. J. Mol. Sci. 24 (2023) 12916. https://doi.org/10.3390/ijms241612916.

[4] P.G. Lyons, C.A. McEvoy, B. Hayes-Lattin, Sepsis and acute respiratory failure in patients with cancer: how can we improve care and outcomes even further?, Curr. Opin. Crit. Care. (2023). https://doi.org/10.1097/MCC.0000000000001078.

[5] A.A. Bardekar, et al., Early prediction of sepsis using Machine Learning Algorithm: A brief clinical perspective, EPRA Int. J. Multidiscip. Res. (2022) 41-45. https://doi.org/10.36713/epra10142.

[6] C.D. Ahlberg, et al., Linking Sepsis with chronic arterial hypertension, diabetes mellitus, and socioeconomic factors in the United States: A scoping review, J. Crit. Care. (2023). https://doi.org/10.1016/j.jcrc.2023.154324.

[7] R. Haas, S.C. McGill, Artificial Intelligence for the prediction of sepsis in adults, Can. J. Health Technol. 2(3) (2022) 1-7. https://doi.org/10.51731/cjht.2022.2837.

[8] M. Reyna, et al., Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019.

[9] L.M. Fleuren, et al., Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy, Intensive Care Med. 46 (2020) 383–400. https://doi.org/10.1007/s00134-019-05872-y.

[10] O.E. Par, E.A. Sezer, H. Sever, Application of Artificial Intelligence in Early–Stage Diagnosis of Sepsis, in: Proc. 2022 5th Artificial Intelligence and Cloud Computing Conference, Association for Computing Machinery, New York, NY, USA, 2023, pp. 196–206. https://doi.org/10.1145/3582099.3582129.

[11] J. Kim, et al., Predicting Cardiac Arrest and Respiratory Failure Using Feasible Artificial Intelligence with Simple Trajectories of Patient Data, J. Clin. Med. 8 (2019) 1336. https://doi.org/10.3390/jcm8091336.

[12] M.M. Islam, et al., Prediction of sepsis patients using machine learning approach: A meta-analysis, Comput. Methods Programs Biomed. 170 (2019) 1-9. https://doi.org/10.1016/j.cmpb.2018.12.027.

[13] Y. Li, Y. Guo, D. Chen, Emergency mortality of non-trauma patients was predicted by qSOFA score, PLoS ONE (2021). https://doi.org/10.1371/journal.pone.0247577.

[14] A.A. Berger, et al., Icosapent Ethyl (Vascepa®) for the Treatment of Acute, Severe Pancreatitis, Cureus 12(11) (2020). https://doi.org/10.7759/cureus.11551.

[15] A. Rhodes, et al., Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016, Intensive Care Med. 43(3) (2017) 304-377.

[16] I. W. Aryabiantara, M. Wiryana, K. Sinardja, T.G.A. Senapathi, I. Widnyana, P.A.S. Panji, I. Sidemen, A. Pradhana, "Comparative Validity Sequential Scoring System Organ Failure Assessment (SOFA) and Quick - Sequential Organ Failure Assessment (qSOFA) on Estimating Mortality for Patients Treated in the Intensive Care Unit of Sanglah General Hospital," J. Anesth. Clin. Res. 08 (2017) 10.4172/2155-6148.1000726.

[17] C. Kok, J. Vicnesh, O. Shu Lih, Z. Xujuan, G. Raj, T. Xiaohui, C. Kang Hao, G. Rashmi, M. Filippo, A.U. Rajendra, "Automated prediction of sepsis using temporal convolutional network," Comput. Biol. Med. 127 (2020) 103957, https://doi.org/10.1016/j.compbiomed.2020.103957.

[18] B. Bataille, J. de Selle, P.E. Moussot, P. Marty, S. Silva, P. Cocquet, "Machine learning methods to improve bedside fluid responsiveness prediction in severe sepsis or septic shock: an observational study," Br. J. Anaesth. 126 (4) (2021) 826-834, https://doi.org/10.1016/j.bja.2020.11.039.

[19] J.S. Calvert, D.A. Price, U.K. Chettipally, C.W. Barton, M.D. Feldman, J.L. Hoffman, M. Jay, R. Das, "A computational approach to early sepsis detection," Comput. Biol. Med. 74 (2016) 69-73, https://doi.org/10.1016/j.compbiomed.2016.05.003.

[20] H.J. Kam, H.Y. Kim, “Learning representations for the early detection of sepsis with deep neural networks,” Comput. Biol. Med. 89 (2017) 248-255, https://doi.org/10.1016/j.compbiomed.2017.08.015.

[21] M. Scherpf, F. Gräßer, H. Malberg, S. Zaunseder, “Predicting sepsis with a recurrent neural network using the MIMIC III database,” Comput. Biol. Med. 113 (2019) 103395, https://doi.org/10.1016/j.compbiomed.2019.103395.

[22] Z. Liu, A. Khojandi, A. Mohammed, X. Li, L.K. Chinthala, R.L. Davis, R. Kamaleswaran, “HeMA: A hierarchically enriched machine learning approach for managing false alarms in real time: A sepsis prediction case study,” Comput. Biol. Med. 131 (2021) 104255, https://doi.org/10.1016/j.compbiomed.2021.104255.

[23] N. Nesaragi, S. Patidar, V. Aggarwal, “Tensor learning of pointwise mutual information from EHR data for early prediction of sepsis,” Comput. Biol. Med. 134 (2021) 104430, https://doi.org/10.1016/j.compbiomed.2021.104430.

[24] Md. M. Islam, T. Nasrin, B.A. Walther, C. Wu, H. Yang, Y. Li, “Prediction of sepsis patients using machine learning approach: A meta-analysis,” Comput. Methods Programs Biomed. 170 (2019) 1-9, https://doi.org/10.1016/j.cmpb.2018.12.027.

[25] N. Kijpaisalratana, D. Sanglertsinlapachai, S. Techaratsami, K. Musikatavorn, J. Saoraya, “Machine learning algorithms for early sepsis detection in the emergency department: A retrospective study,” Int. J. Med. Inform. 160 (2022) 104689, https://doi.org/10.1016/j.ijmedinf.2022.104689.

[26] T. Aşuroğlu, H. Oğul, “A deep learning approach for sepsis monitoring via severity score estimation,” Comput. Methods Programs Biomed. 198 (2021) 105816, https://doi.org/10.1016/j.cmpb.2020.105816.

[27] R.J. Delahanty, J. Alvarez, L.M. Flynn, R.L. Sherwin, S.S. Jones, “Development and Evaluation of a Machine Learning Model for the Early Identification of Patients at Risk for Sepsis,” Ann. Emerg. Med. 73 (4) (2019) 334-344, https://doi.org/10.1016/j.annemergmed.2018.11.036.

[28] L. Zhang, Z. Wang, Z. Zhou, S. Li, T. Huang, H. Yin, J. Lyu, “Developing an ensemble machine learning model for early prediction of sepsis-associated acute kidney injury,” iScience 25 (9) (2022) 104932, https://doi.org/10.1016/j.isci.2022.104932.

[29] N. Ghias, S. Ul Haq, H. Arshad, H. Sultan, F. Bashir, S. A. Ghaznavi, M. Shabbir, Y. Badshah, M. Rafiq, “Using Machine Learning Algorithms to predict sepsis and its stages in ICU patients,” medRxiv (2022) https://doi.org/10.1101/2022.03.15.22271655.

[30] M. Scherpf, F. Gräßer, H. Malberg, S. Zaunseder, “Predicting sepsis with a recurrent neural network using the MIMIC III database,” Comput. Biol. Med. 113 (2019) 103395, https://doi.org/10.1016/j.compbiomed.2019.103395.

[31] E. Bloch et al., "Machine Learning Models for Analysis of Vital Signs Dynamics: a Case for Sepsis Onset Prediction," Journal of Healthcare Engineering, vol. 2019, p. 5930379, 2019.

[32] X. Zhao, W. Shen, G. Wang, "Early Prediction of Sepsis Based on Machine Learning Algorithm," Computational Intelligence and Neuroscience, vol. 2021, Article ID 6522633, 13 pages, 2021. [Online]. Available: https://doi.org/10.1155/2021/6522633

[33] Y. V. Singh, P. Singh, S. Khan, R. S. Singh, "A Machine Learning Model for Early Prediction and Detection of Sepsis in Intensive Care Unit Patients," Journal of Healthcare Engineering, vol. 2022, Article ID 9263391, 11 pages, 2022. [Online]. Available: https://doi.org/10.1155/2022/9263391

[34] E. Persad, K. Jost, A. Honoré, D. Forsberg, K. Coste, H. Olsson, S. Rautiainen, E. Herlenius, "Neonatal sepsis prediction through clinical decision support algorithms: A systematic review," Acta Paediatrica, vol. 110, 2021. doi: 10.1111/apa.16083.

[35] A. Honoré, D. Forsberg, K. Adolphson, S. Chatterjee, K. Jost, E. Herlenius, "Vital sign-based detection of sepsis in neonates using machine learning," Acta Paediatr., Jan. 2023. [Epub ahead of print]. doi: 10.1111/apa.16660.

[36] A.J. Masino, M.C. Harris, D. Forsyth, S. Ostapenko, L. Srinivasan, C.P. Bonafide, et al., "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," PLoS ONE, vol. 14(2), e0212665, 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0212665

[37] S. Le, J. Hoffman, C. Barton, J.C. Fitzgerald, A. Allen, E. Pellegrini, J. Calvert, R. Das, "Pediatric Severe Sepsis Prediction Using Machine Learning," Frontiers in Pediatrics, vol. 7, 2019. doi: 10.3389/fped.2019.00413.

[38] J.E. Camacho-Cogollo, I. Bonet, B. Gil, E. Iadanza, "Machine Learning Models for Early Prediction of Sepsis on Large Healthcare Datasets," Electronics, vol. 11, 1507, 2022. [Online]. Available: https://doi.org/10.3390/electronics11091507

[39] D. Wang, J. Li, Y. Sun, X. Ding, X. Zhang, S. Liu, B. Han, H. Wang, X. Duan, T. Sun, "A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients," Frontiers in Public Health, vol. 9, 2021. doi: 10.3389/fpubh.2021.754348.

[40] C.J. Chiew et al., "Heart rate variability based machine learning models for risk prediction of suspected sepsis patients in the emergency department," Medicine, vol. 98(6), p e14197, Feb. 2019. doi: 10.1097/MD.0000000000014197.

[41] K. Li, Q. Shi, S. Liu, Y. Xie, J. Liu, "Predicting in-hospital mortality in ICU patients with sepsis using gradient boosting decision tree," Medicine, vol. 100(19), p e25813, May 2021. doi: 10.1097/MD.0000000000025813.

[42] M. Komorowski, A. Green, K.C. Tatham, C. Seymour, D. Antcliffe, "Sepsis biomarkers and diagnostic tools with a focus on machine learning," EBioMedicine, vol. 86, 104394, Dec. 2022. [Online]. Available: https://doi.org/10.1016/j.ebiom.2022.104394

[43] H.-F. Deng et al., "Evaluating machine learning models for sepsis prediction: A systematic review of methodologies," iScience, vol. 25, issue 1, 103651, 2022. [Online]. Available: https://doi.org/10.1016/j.isci.2021.103651.

[44] S.L. Kausch, J.R. Moorman, D.E. Lake, J. Keim-Malpass, "Physiological machine learning models for prediction of sepsis in hospitalized adults: An integrative review," Intensive and Critical Care Nursing, vol. 65, 103035, 2021. [Online]. Available: https://doi.org/10.1016/j.iccn.2021.103035.

[45] S. Raza, "Improving Clinical Decision Making with a Two-Stage Recommender System: A Case Study on MIMIC-III Dataset," medRxiv, 2023. [Online]. Available: https://doi.org/10.1101/2023.02.21.23286247.

[46] P. Marik, A. Taeb, "SIRS, qSOFA and new sepsis definition," Journal of Thoracic Disease, vol. 9, no. 4, pp. 943-945, 2017. doi: 10.21037/jtd.2017.03.125.

[47] R. Moreno, A. Rhodes, L. Piquilloud, et al., "The Sequential Organ Failure Assessment (SOFA) Score: has the time come for an update?," Critical Care, vol. 27, p. 15, 2023. doi: 10.1186/s13054-022-04290-9.

[48] B. Vidal, S. Fieux, M. Colom, et al., "18F-F13640 preclinical evaluation in rodent, cat and primate as a 5-HT1A receptor agonist for PET neuroimaging," Brain Struct Funct, vol. 223, pp. 2973–2988, 2018. [Online]. Available: https://doi.org/10.1007/s00429-018-1672-7.

[49] M.E. Nunnally, R. Ferrer, G.S. Martin, et al., "The Surviving Sepsis Campaign: research priorities for the administration, epidemiology, scoring and identification of sepsis," ICMx, vol. 9, p. 34, 2021. [Online]. Available: https://doi.org/10.1186/s40635-021-00400-z.

[50] F. Gül, M.K. Arslantaş, İ. Cinel, A. Kumar, "Changing Definitions of Sepsis," Turkish Journal of Anaesthesiology and Reanimation, vol. 45, no. 3, pp. 129-138, Jun. 2017. doi: 10.5152/TJAR.2017.93753. PMID: 28752002; PMCID: PMC5512390.

[51] N.M. Elmahdy, T. Elsenousy, D.A. Abdelatief, D.M. Maarouf, "Prognostic Scoring Systems as a Tool to Predict the Clinical Outcomes for Patient with Critical Condition," Egyptian Journal of Health Care, vol. 13, no. 4, pp. 1385-1402, 2022. doi: 10.21608/ejhc.2022.270426.

[52] J.B. Lascarrou, H. Merdji, A. Le Gouge, et al., "Targeted Temperature Management for Cardiac Arrest with Nonshockable Rhythm," N Engl J Med, vol. 381, pp. 2327-2337, Dec. 2019. doi: 10.1056/NEJMoa1906661.

[53] L. Calandriello, E. De Lorenzis, G. Cicchetti, et al., "Extension of Lung Damage at Chest Computed Tomography in Severely Ill COVID-19 Patients Treated with Interleukin-6 Receptor Blockers Correlates with Inflammatory Cytokines Production and Prognosis," Tomography, vol. 9, pp. 981-994, 2023. [Online]. Available: https://doi.org/10.3390/tomography9030080.

[54] A.T. Yacoub, L. Mojica, L. Jones, A. Knab, S. Alrabaa, J. Greene, "The Role of Corticosteroids in Adult Respiratory Distress Syndrome caused by Viridans Group Streptococci Bacteremia in Neutropenic Patients," DH. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, unpublished.

[55] [Online]. Available: https://www.scribd.com/document/330741997/Clinical-Intensive-Care-and-Acute-Medicine-2nd-Ed-pdf. "Clinical Intensive Care and Acute Medicine, 2nd Ed.pdf." (Access Date: August, 2023)

[56] A.D. Shields, L.A. Plante, L.D. Pacheco, J.M. Louis, "Society for Maternal-Fetal Medicine Consult Series #67: Maternal sepsis," Society for Maternal-Fetal Medicine (SMFM), May 24, 2023. [Online]. Available: https://doi.org/10.1016/j.ajog.2023.05.019.

[57] [Online]. Available: https://vdoc.pub/documents/predictive-analytics-and-data-mining-concepts-and-practice-with-rapidminer-4dlmce9stuq0. "Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner."

[58] H.D. Panchal, H.B. Shah, "Multiple forgery detection in digital video based on inconsistency in video quality assessment attributes," Multimedia Systems, vol. 29, pp. 2439–2454, 2023. [Online]. Available: https://doi.org/10.1007/s00530-023-01123-9.

[59] [Online]. Available: https://www.science.gov/topicpages/r/repeated+head+trauma. "Topic: repeated head trauma." (Access Date: August, 2023)

[60] S. Catalano, J. Moyer, A. Weaver, Q. Di, J.D. Schwartz, M. Catalano, C.K. Ward-Caviness, "Associations between long-term fine particulate matter exposure and hospital procedures in heart failure patients," unpublished.

[61] [Online]. Available: https://www.coursehero.com/file/118380998/Sensitivity-and-Specificitydocx/. "Sensitivity and Specificity."(Access Date: August, 2023)

[62] Y. Matsuzaka, Y. Uesawa, "Optimization of a Deep-Learning Method Based on the Classification of Images Generated by Parameterized Deep Snap a Novel Molecular-Image-Input Technique for Quantitative Structure–Activity Relationship (QSAR) Analysis," Frontiers in Bioengineering and Biotechnology, vol. 7, 2019. doi: 10.3389/fbioe.2019.00065.

[63] [Online]. Available: https://mohcinemadkour.github.io/posts/2018/06/Analysing%20Model%20Perfromance%20from%20ROC,%20and%20Recall%20and%20Precision%20curves/. "Analysing Model Performance from ROC, and Recall and Precision curves." (Access Date: August, 2023)

[64] A. Salam, F. Ullah, F. Amin, M. Abrar, "Deep Learning Techniques for Web-Based Attack Detection in Industry 5.0: A Novel Approach," Technologies, vol. 11, p. 107, 2023. doi: 10.3390/technologies11040107.

[65] F. Zabihollahy, N. Schieda, E. Ukwatta, "Patch-Based Convolutional Neural Network for Differentiation of Cyst From Solid Renal Mass on Contrast-Enhanced Computed Tomography Images," IEEE Access, vol. 8, 2020. doi: 10.1109/ACCESS.2020.2964755.

[66] P. Santhiya, S. Kavitha, T. Aravindh, S. Archana, A.V. Praveen, "Fake News Detection Using Machine Learning," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-8. doi: 10.1109/ICCCI56745.2023.10128339.

[67] H. Lassoued, R. Ketata, S. Yacoub, "ECG Decision Support System based on feedforward Neural Networks," International Journal on Smart Sensing and Intelligent Systems, vol. 11, no. 1, pp. 1-15, 2018. doi: 10.21307/ijssis-2018-029.

[68] S.R. Cheers, A.E. O'Connor, T.K. Johnson, D.J. Merriner, M.K. O'Bryan, J.E.M. Dunleavy, "Spastin is an essential regulator of male meiosis, acrosome formation, manchette structure, and nuclear integrity," bioRxiv, 2022. doi: 10.1101/2022.08.01.502419.

[69] V.K.N.K. Pai, M. Balrai, S. Mogaveera, D. Aeloor, "Face Recognition Using Convolutional Neural Networks," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2018, pp. 165-170. doi: 10.1109/ICOEI.2018.8553969.

[70] V.S. Gaur, V. Sharma, J. McAllister, "Abusive adversarial agents and attack strategies in cyber-physical systems," CAAI Trans. Intell. Technol., vol. 8, no. 1, pp. 149–165, 2023. doi: 10.1049/cit2.12171.

[71] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

[72] M.L. Minsky, S.A. Papert, Perceptrons. Cambridge, MA: MIT Press, 1969.

[73] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning. MIT Press, 2016.

[74] W.S. McCulloch, W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bulletin of Mathematical Biophysics, vol. 5, pp. 115–133, 1943.

[75] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[76] S. Pandey, P. Dadure, M.V.L. Nunsanga, P. Pakray, "Parts of speech tagging towards classical to quantum computing," 2022 IEEE Silchar Subsection Conference (SILCON), Silchar, India, 2022, pp. 1-6. doi: 10.1109/SILCON55242.2022.10028796.

[77] J. Yin, Z. Deng, A.V.M. Ines, J. Wu, E. Rasu, "Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (Bi-LSTM)," Agricultural Water Management, vol. 242, 2020, 106386. doi: 10.1016/j.agwat.2020.106386.

[78] H. Lyu, H. Lu, L. Mou, "Learning a Transferable Change Rule from a Recurrent Neural Network for Land Cover Change Detection," Remote Sens., vol. 8, 506, 2016. doi: 10.3390/rs8060506.

[79] J. Cui, W. Kong, X. Zhang, D. Chen, Q. Zeng, "DLSTM-Based Successive Cancellation Flipping Decoder for Short Polar Codes," Entropy, vol. 23, 863, 2021. doi: 10.3390/e23070863.

[80] A. Ye, Z. Wang, Modern Deep Learning for Tabular Data, Novel Approaches to Common Modeling Problems. Apress, 2023.

[81] Y. Chen, J. Yang, K. Zhang, Y. Xu, Y. Liu, "A Feature-Cascaded Correntropy LSTM for Tourists Prediction," IEEE Access, vol. 9, pp. 32810-32822, 2021. doi: 10.1109/ACCESS.2021.3059943.

[82] D.L.X. Fung et al., "A self-knowledge distillation-driven CNN-LSTM model for predicting disease outcomes using longitudinal microbiome data," Bioinformatics Advances, vol. 3, no. 1, 2023, vbad059. doi: 10.1093/bioadv/vbad059.

[83] Y. Peng, M. Liao, H. Deng, L. Ao, Y. Song, W. Huang, J. Hua, "CNN–SVM: a classification method for fruit fly image with the complex background," IET Cyber-Physical Systems: Theory & Applications, 2020. doi: 10.1049/iet-cps.2019.0069.

[84] M. Desai, M. Shah, "An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)," Clinical eHealth, vol. 4, pp. 1-11, 2021. doi: 10.1016/j.ceh.2020.11.002.

[85] [Online]. Available: https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939. "Convolutional Neural Networks Explained." (Access Date: September, 2023)

[86] F. Stuker, J. Ripoll, M. Rudin, "Fluorescence molecular tomography: Principles and potential for pharmaceutical research," Zurich Open Repository and Archive, University of Zurich, 2023.

[87] L. Buturović, D. Miljković, "A novel method for classification of tabular data using convolutional neural networks," bioRxiv, 2020. doi: 10.1101/2020.05.02.074203.

[88] C.-C. J. Kuo, "Understanding convolutional neural networks with a mathematical model," Journal of Visual Communication and Image Representation, vol. 41, pp. 406-413, 2016. doi: 10.1016/j.jvcir.2016.11.003.

[89] [Online]. Available: https://www.ibm.com/topics/random-forest. "What is Random Forest." (Access Date: September, 2023)

[90] N.E. Rodríguez-Maya, J.J. Flores, S. Verel, et al., "Models to classify the difficulty of genetic algorithms to solve continuous optimization problems," Natural Computing, 2023. doi: 10.1007/s11047-022-09936-9.

[91] V.V. Prasad et al., "Prediction of Stock Prices Using Statistical and Machine Learning Models: A Comparative Analysis," The Computer Journal, vol. 65, no. 5, pp. 1338–1351, 2022. doi: 10.1093/comjnl/bxab008.

[92] P.-Y. Yin, C.-C. Tsai, R.-F. Day, "PSO active learning of XGBoost and spatiotemporal data for PM2.5 sensor calibration," IOP Conference Series: Earth and Environmental Science, vol. 227, no. 5, 2019. doi: 10.1088/1755-1315/227/5/052048.

[93] "Development and Comparison of Virtual Sensors Constructed using AI Techniques to Estimate the Performances of IC Engines," SAE Technical Paper 2022-01-1064, 2022. doi: 10.4271/2022-01-1064.

APPENDICES

Appendix 1 - Literature review

Category	Year / Ref	Aim	Attributes	Data Source and Points	Criterias	Missing Values	Sepsis Definition	Methods	Metric
Diagnosis	2021 - [18]	To evaluate the effectiveness of machine learning techniques in using transthoracic echocardiography (TTE) data to predict fluid responsiveness in critically ill patients.	Thoracic ultrasonographic data	*Narbonne Hospital *100 patients	Sequential Organ Failure Assessment Score >=2	Non-linear iterative PLS algorithm	SOFA	CART (classification and regression tree) PLS (partial least-squares regression) NNET LDA (linear discriminant analysis)	AUCs PLR 0.77 , CART 0.68, PLS 0.83 (clinically meaningful), NNET 0.83, LDA 0.85
Early Diagnosis	2016 - [19]	To predict the onset of sepsis three hours before its initial sustained episode in adults newly admitted to the ICU who are not already suffering from sepsis.	Systolic blood pressure,pulse pressure, hear trate, temperature, respiration rate, white blood cell count, pH, blood oxygen saturation and age	* MIMICII *1394 patients	*Adult patient *Patient does not meet SIRS criteria at time of admission *Documented measurements available for attributues.	Most recent available observation	SIRS	Regression	Sensitivity: 0.90 Specificity: 0.81 AUC : 0.83
Early Diagnosis	2016 - [20]	To build computational models using deep learning to identify the initial stages of sepsis within a specific five-hour window.	Systolic blood pressure, pulse pressure, heart rate, body temperature, respiration rate, white blood cell count, pH, blood oxygen saturation, and age.	* MIMICII * 5803 patients	*Adult patient * Patients who fulfilled the criteria for SIRS during (a) the first 4 h of hospital admission or (b) the first 1 h of the first ICU admission.	The nearest measured value	SIRS	MLP	Accuracy : 0.915 Sensitivity: 0.886 Specificity: 0.944 AUROC : 0.915
Early Diagnosis	2019 - [21]	To use a Recurrent Neural Network to predict the onset of sepsis three hours	Age, systolic blood pressure, diastolic blood pressure, pH value, blood oxygen saturation, temperature, respiration rate, white blood cell count, CO2 partial pressure	* MIMICIII * 32520 patients	*Adult patient * At least one measurement for each of the SIRS parameters * Exclusuion of addmision with sepsis diagnosis and onset before the sum of predicition time and a minimal 5 hours of duration	Last observation carried forward	SIRS	RNN	AUC : 0.81 sensitivity : 0.90 Specificty: 0.47

Early Diagnosis	2021 - [22]	To develop a two-phase architecture called HeMA that uses machine learning algorithms and statistical evaluations.	Heart Rate, blood pressures, respiration rate and oxygen saturation.	* Cerner CareAware iBus * 634 patients	NA	Last observation carried forward	SOFA	RF NN	RF Sensitivity: 0.92 Specificity: 0.37 AUROC : 0.72 Lead time (min): 332 NN Sensitivity: 0.92 Specificity: 0.22 AUROC : 0.64 Lead time (min): 343
Early Diagnosis	2021 - [23]	To create a machine learning-based framework that uses physiological data from electronic health records to predict the onset of sepsis up to six hours in advance.	Heart rate, O2Sat, Temperature, SBP, MAP, DBP, Resp, Bicarbonate, FiO2, Ph, PaCO2, BUN, Calcium, Chloride, Creatinine, Gl, Magnesium, Phosphate, Potassium, Hematocrit, Hemoglobin, PTT, WBC, Platelet count	* PhysioNet * 40336 ICU patients	Having missing data less than 95% are only considered.	Forward fill	SOFA	Gradient boosting machine learning models	AUC: 0.8621 F1: 0.1636
Diagnosis	2022 - [25]	To develop and assess the innovative sepsis diagnostic tools employing machine learning algorithms.	Temperature, systolic and diastolic blood pressure, pulse, respiratory rate, age, sex, mode of arrival, emergency severity index, comorbidities and free text (from EHR)	* Faculty of Medicine, Chulalongkorn University (Thai) *133707 patients	*Adult patient	Means and medians	qSOFA ≥ 2, MEWS ≥ 5 and SIRS ≥ 2	Logistic Regression, Gradient Boosting, Random Forest, Neural network	Logistic Regression AUC : 0,930 Sensitivity : 0,8694 Specificity : 0,8725 Gradient Boosting AUC :0,919 Sensitivity : 0,8449 Specificity : 0,8480 Random Forest AUC : 0,931 Sensitivity : 0,8694 Specificity : 0,8692 Neural Network AUC : 0,926 Sensitivity : 0,8571 Specificity : 0,8631
Early Diagnosis	2021 - [26]	To predict the SOFA score for patients up to six hours before the onset of sepsis using only seven essential physiological metrics easily obtained at an ICU bedside.	Heart rate, respiratory rate, systolic blood pressure,diastolic blood pressure, oxygen saturation, temperature , glascow comma score	*MIMIC III * 5154	*at least one measurement is available for each of the vital signs 12 hour prior to predicted onset time.	Probabilistic Principal Component Analysis (PPCA) method	SOFA	CNN and RF (Ensemble method)	AUC : 0,982 TPR : 0,984 FPR : 0,032 FNR: 0,016 TNR: 0,968
Diagnosis	2019 - [27]	To create a new sepsis detection tool called the Risk of Sepsis (RoS) score using machine learning algorithms and compares its effectiveness with existing sepsis metrics like SIRS, SOFA, qSOFA, MEWS, and NEWS.	Subset of laboratory results, vital signs, demographics, administered medications, nursing documentation, and key words extracted from the ED chief complaint.	* Urban Community Hospital * 2759529	*Adult patient * Lab results or vital signs were documented	NA	SOFA, qsofa, MEWS, NEWS	Gradient boosting	AUC: 0.97 Senstivitiy: 0.677 Specificity: 0.964

Early Diagnosis	2019 - [3.4]	To develop a machine learning algorithm for the early detection of sepsis by analyzing physiological data from 1,161 critically ill patients, with the model successfully predicting sepsis about 5 hours before its actual onset.	Heart rate, diastolic and systolic blood pressure, mean arterial pressure, temperature, respiratory rate, peripheral oxygen saturation, and white blood cell count.	* Methodist LeBonheur Healthcare (MLH) System in Memphis * 1161 patients	*Adult patient * Patients with continous physical data * Patients without cardiovascular disease	NA	SIRS	RF, SVM, LR, MLP and RNN.	F1 score of up to 80% and 67% one hour before sepsis onset. On average, these models were able to predict sepsis 294.19 ± 6.50 min (5 h) before the onset.
Diagnosis	2022 - [3.5]	To examine the negative impact of data drift on the effectiveness of machine learning algorithms for predicting sepsis	age, gender, Systolic Arterial Blood Pressure (SysAPB), Diastolic Arterial Blood Pressure (DiasABP), Heart Rate (HR), Respiration Rate (RespRate), and Oxygen Saturation (SpO2).	* 4 EDs from USA * 112972 patients	*Adult patient	NA	SOFA	XGBoost RNN (GRU)	Major event simulation (AUC) XGBoost : 0.866 Covariate Shift XGBoost : 0.873 Concept shift XGBoost : 0.842
Diagnosis	2022 - [28]	To develop an accurate analytical model for predicting the risk of sepsis-associated acute kidney injury using easily available clinical data.	Heart rate, creatinine, temperature, PaO2, hemoglobin, lactate	* MIMIC-IV ZG Database e-ICU CRD * 21038 patients	Exclusion Criteria * AKI before sepsis * ICU stay <48 hours	Removed	SOFA	Support Vector Machine, Random Forest, Neural Network, and Extreme Gradient Boost Ensemble	Ensemble AUC : 0.756–0.813
Diagnosis	2022 - [29]	To use evidence-based guidelines as a structured framework for managing newborns at risk of early onset sepsis, while also reducing antibiotic usage.	HR, temp, O2Sat, SBP, and MAP at 6hr before prediction	* PhysioNet * 1552210 patients	Variables with more than 70% missing values were dropped	Missforest	NA	Xgboost, Random Forest, and linear learner	XGBoost Accuracy: 0.98 , precision : 0.97, Recall 0.98 AUC : 0.95 to 0.98
Diagnosis	2019 - [30]	To use temporal convolutional network architecture for predicting the onset of sepsis in ICU (Intensive Care Unit) patients who were not initially classified as septic.	Patient age, systolic blood presure , diastolic blood pressure, blood oxygen saturation(SO2), temperature, heart rate, respiratory rate, CO2 partial pressure(PaCO2)	* MIMIC III * NA	*Adult patient	NA	SOFA	Temporal Convolutional Network	Accuracy: 0.785 Precision: 0.829 Recall: 0.619 AUC: 0.770 F1: 0.700
Early Diagnosis	2019 - [31]	To estimate the likelihood of a patient developing sepsis in the next four hours based on data from the previous eight hours.	Heart rate, temperature, mean arterial blood pressure, and respiratory rate	* Rabin Medical Center (RMC) * 600 patients	*Adult patient * Patients stayed a minimum of 12 hours in the ICU * Patients did not meet SIRS criteria at time of admission to the ICU * Continuous documented measurements were available for at least 12 hours for vital signs	NA	SIRS	Support Vector Machine (radial basis function)	AUC: 0.8838 Accuracy: 0.8710 Sensitivity: 0.7792 Specificity: 0.9615

Early Diagnosis	2021 - [32]	To use of specific machine learning methods (XGBoost and LightGBM) to predict the onset of sepsis six hours before it actually occurs.	HR, O2Sat, Temp, SBP, MAP, DBP, Resp, HCO3, pH, PaCO2, AST, BUN, AlkalinePhos, Chloride, Creatinine, Lactate, Magnesium, Potassium, Bilirubin_total, PTT, WBC, Fibrinogen, Platelets, Age, Gender	* Physionet * 22336 patients	Variables with more than 98% missing proportions were removed.	Miceforest	qSOFA	XGBoost LightGBM	XGBoost AUC: 0.9389 Precision: 0.78 Recall: 0.55
Diagnosis	2022 - [33]	To present a machine learning approach for early identification and prediction of sepsis in ICU patients, using clinical laboratory metrics and vital sign data.	Age, heart rate, blood pressure, RR, procalcitonin, temperature, platelet count, WBC count, SOFA score, SBP, DBP, mean arterial pressure (MAP: it is a combination of SBP and DBP), SIRS criteria, oxygen saturation (SpO2), hemoglobin, partial pressure of oxygen (PaO2), creatinine, total time in hospital, Glasgow coma scale (GCS), c-reactive protein serum, and p-lactate	* Skaraborg Hospital * 1572 patients	NA	Average mean	SIRS	Ensemble method (which combines SVM, RF, NB, LR, and XGBoost)	Accuracy : 0,96 Precision :0,98 Recall : 0,94 Specificity : 0,97 F1 score : 0,98 AUC : 0,96
Early Diagnosis	2023 - [35]	To discuss the use of machine learning, specifically a Naïve Bayes algorithm, for predicting sepsis in newborns up to 24 hours before clinical suspicion arises.	inter-beat interval, respiratory rate, and SpO2, postnatal age, birth weight, and sex.	* Karolinska University Hospital Solna and Huddinge * 325 patients (infants)	NA	Linearly interpolated	Nsofa	Naïve Bayes	AUC: 0.82
Diagnosis	2022 - [38]	To introduce a specific methodology for detecting the onset of sepsis in adult ICU patients. This methodology aligns with the SEPSIS-3 criteria, which focus on identifying organ dysfunction and suspected infection. The framework uses 24-hour retrospective data to make predictions in one-hour forecast intervals.	physiological data (e.g., heart rate, temperature, etc.), laboratory test results (e.g., white blood count, glucose, hematocrit, hemoglobin, creatinine, bicarbonate, PH, and arterial blood gases) and demographics/score (age, Elixhauser Index, weight, and Glasgow coma score)	* MIMIC III * 2377 patients	Exclusion Criteria Patient aged <=14, admission cardio surgey, admission with missing data Sepsis pre-ICU intime	k-nearest neighbors imputation	SOFA	RF + SVM ANN + AdaBoost RF + ANN SVM + ANN SVM + ADA + ANN + KNN + RF + Tree ANN + KNN + RF + SVM + Tree XGBoost	XGBoost - Information Gain > 0.002 AUC-ROC : 0.918 Accuracy : 0.872 Recall : 0.852 Precision : 0.868
Early Diagnosis	2019 - [36]	To create a predictive model using electronic health record (EHR) data that can identify the onset of sepsis in infants at least four hours before it is clinically detected.	The authors identified 30 features associated with infant sepsis through literature review and expert consultation.	* Children's Hospital of Philadelphia * 1188 patients	Exclusion Criteria Evaluations within 48 hours Fungal and viral patohens Bacterial patogens identified but not present in blood	Mean imputation	NA	Logistic regression, Naïve Bayes, Support vector machine (SVM), K-nearest neighbors (KNN), Random forest, AdaBoost, and Gradient boosting.	SVM AUC : 0.86 Sensitivity: 0.74 Specificity0.79
Early Diagnosis	2019 - [37]	To evaluate the effectiveness of a machine-learning algorithm in predicting the onset of severe sepsis in pediatric groups about four hours before it occurs, using electronic healthcare record (EHR) data.	Patient age, diastolic and systolic blood pressures, heart rate, temperature, respiration rate, and peripheral oxygen saturation	* University of California San Francisco (UCSF) Medical Center * 4449	Exclusion Criteria <6 hours after the start of the patient record	Removed	SIRS	MLA	AUC: 0.916 Sensitiviy: 0.750 Specificity: 0.940

Diagnosis	2021 - [39]	To develop an artificial intelligence algorithm aimed at early prediction of sepsis, in order to address a significant ongoing challenge that affects patients, healthcare professionals, and medical infrastructures globally.	Nutrophils, D-Dimer, Albunim, WBC, Direct Biliribun, Potassium, Calcium, Magnessium, LDL, Uric acid, sex, age...etc. Totally 20 attribitues	* ICU of the First Affiliated Hospital of Zhengzhou University * 4449	Exclusion criteria Age <18 Patients with heart disease, fracture, neoplasm, cerebral infaction More than 3 missing data	NA	SIRS	Random Forest	AUC: 0.91, Sensitivity: 0.87, Specificity: 0.89.
Mortality	2019 - [40]	To compare the effectiveness of machine learning models based on heart rate variability (HRV) with existing risk assessment tools like qSOFA, NEWS, MEWS, and the Singapore ED Sepsis (SEDS) model. The aim is to improve the accuracy of predicting 30-day in-hospital mortality rates for patients suspected of having sepsis in emergency departments.	age, ethnicity, gender, temperature, heart rate, respiratory rate, systolic bp, diastolic bp, GCS	* Singapore General Hospital ED * 214 patients	Patients above 21 years and who met at least 2 of 4 SIRS criteria	NA	SIRS	k-nearest neighbors, random forest, adaptive boosting, gradient boosting, and support vector machine.	gradient boosting, with an F1 score of 0.50 and an area under the precision-recall curve (AUPRC) of 0.35.
Mortality	2021 - [41]	To establish a clinical decision support mechanism specifically aimed at predicting sepsis-related mortality in ICU settings.	Age, sex, ethnicity, lenght of hospital stay, Glasgow Coma Sclae, oxygen saturation, vital signs, labarotary values	* MIMIC-III * 3937 patients	Patients aged 18 years or older and excluded patients with data missing more than 30%	Mean of each group	ICD-9	GBDT, LR, KNN, RF, and SVM	GBDT AUROC (0.992), highest precision (94.8%), recall (91.7%), accuracy (95.4%), and F1 score (0.933)
Diagnosis	2019 - [42]	To evaluate the impact of this computational tool on clinical practices and patient outcomes. The ultimate goal is to address the limitations of existing methods for sepsis prediction and improve the accuracy of sepsis identification and forecasting in non-ICU inpatient settings.	Age, Blood pressure, heart rate, blood urea nitogen, creatine, Creatinine, Absolute Lymphocyte Count	* Tertiary teaching hospital system in Philadelphia * 162212 patients	NA	NA	SIRS	RF	a sensitivity of 26% and specificity of 98%, with a positive predictive value of 29% and positive likelihood ratio of 13.

Appendix 2 - Publications

1) Oznur Esra Par, Ergin Sosyal, “A Clinical Decision Support System Model for Cardiovascular Diseases”, IX. National Medical Informatics Congree, Turkey, 2012

2) Oznur Esra Par, Ergin Sosyal, “Evaluation of Clinical Decision Support Systems”, XXIX. National Informatics Congree, Turkey, 2012

3) Oznur Esra Par, Ebru Akcapinar Sezer, Hayri Sever, “Small and Unbalanced Data Set Problem in Classification”, 27th Signal Processing and Communications Applications Conference, SIU 2019, Sivas, Turkey, 2019

4) Oznur Esra Par, Ebru Akcapinar Sezer, Hayri Sever, ”Clinical Decision Support Systems: From the Perspective of Small and Imbalanced Data Set,” 17th International Conference on Informatics, Management, and Technology in Healthcare (ICIMTH) , vol.262, Athens, Greece, pp.344-347, 2019

5) Oznur Esra Par, Ebru Sezer and Hayri Sever, “Application of Artificial Intelligence in Early–Stage Diagnosis of Sepsis”, AICCC 2022, Osaka, Japan, December 17-19, 2022. (The best presentation of Session 6)