History of Medicine, 2023, 9(2):
DOI: 10.48047/HM.V9.I2.2023.

Applying Advanced Computational Models to Optimise Electronic Health Records and Enhance Clinical Decision-Making

Busayo Jegede[1*]

[1]Doctorate of Business administration, Department of Healthcare administration, Indiana Wesleyan university, Indiana, United states

Corresponding Author Email: busayojegede1960@gmail.com

Abstract:
Electronic Health Records (EHRs) have become a cornerstone in modern healthcare, providing structured patient data that can significantly enhance clinical decision-making. However, the manual interpretation of large-scale EHR data presents challenges in efficiency and accuracy, necessitating the integration of advanced computational models. This study explores machine learning and deep learning techniques to optimise EHR-based cardiovascular disease (CVD) prediction, leveraging data-driven insights to enhance patient risk stratification. The Cardiovascular Disease Dataset from Kaggle, comprising 70,000 patient records, was utilised to train and evaluate three models: Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks. Among the three models, XGBoost demonstrated the highest predictive accuracy at 73.5%, making it the most effective model for CVD detection. LSTM exhibited the highest recall (0.81), making it well-suited for identifying high-risk patients, but it also generated a higher number of false positives, potentially leading to unnecessary medical interventions. Random Forest, a baseline model, achieved 71.2% accuracy, showing stable but slightly lower performance. These findings highlight the superiority of XGBoost in predictive accuracy, while LSTM remains useful in maximising sensitivity to disease detection. The results emphasise the potential of machine learning in automated cardiovascular risk assessment, allowing for data-driven clinical decision-making that can enhance early intervention strategies. The study's implications extend to EHR optimisation, AI integration in medical workflows, and the deployment of computational models for clinical risk management.

Keywords: Electronic Health Records (EHRs), Machine Learning in Healthcare, Clinical Decision Support Systems (CDSS), Cardiovascular Disease Prediction, Random Forest, XGBoost, LSTM, Predictive Analytics in Medicine.

I. INTRODUCTION

Cardiovascular diseases (CVDs) are one of the top causes of global health problems since the World Health Organization (WHO) reports annual deaths reaching 17.9 million [1]. The expanding CVD prevalence requires the development of early detection systems, which help slow disease advancement and decrease healthcare expenses. Traditional disease diagnostic processes depend on clinical signs assessment, doctor expertise and handling of medical records, yet these approaches slow down prompt disease recognition. Hospital systems implementing Electronic Health Records acquire broad patient databases containing individual information, medical history, and test results supporting evidence-based healthcare decisions [2].

EHRs remain limited by three key obstacles: their disordered structure, comprehensive complexity, and vast information needs for analysis [3]. Physicians find it challenging to identify essential information patterns within EHR datasets, which delays medical diagnosis decisions and thus affects patient care results [4]. ML and DL models enable efficient analysis of big EHR datasets, helping medical staff identify patients at high risk [5]. After proper optimisation, machine learning models improve medical diagnoses by reducing physician mistakes and increasing predictive healthcare capabilities.

The research evaluates machine learning methods for designing CVD prediction systems utilising computer-generated models trained using structured EHR data systems. Random Forest, coupled with XGBoost and LSTM models, is the key method for finding the best clinical decision enhancement solution. XGBoost demonstrates substantial classification accuracy because of ensemble learning processes. At the same time, LSTM achieves its unique expertise in recognising sequential patterns in deep learning systems to extract long-term dependencies from patient histories [6]. Random Forest performs as a benchmark model so physicians can understand the reference point for advanced models.

The recent popularity of EHR-based predictive modelling in healthcare analytics data cannot conceal the present obstacles which limit the performance of EHR-driven disease prediction systems [7]. The system faces three main limitations: it cannot process data immediately, and there are variations in feature weights while understanding medical patterns. Diagnostic methods for CVDs must expand their analysis to include all patient history stored in EHRs because doing so affects accuracy in disease progression forecasting [8]Current machine learning applications in healthcare mainly perform single-model assessments and omit systematic analyses about the effectiveness of tree-based ensemble models (Random Forest, XGBoost) compared to deep learning architectures (LSTM).

A complete assessment framework for model evaluation in EHR-based CVD prediction remains absent from the literature, which hinders the selection process of the most powerful computational methods for automated risk identification and clinical decision-making systems (CDSS) [9]. Real-time hospital deployment of these models faces scalability limitations because most lack the necessary efficiency levels. The research investigates two essential issues: (1) it selects the top predictive model for CVD diagnosis using actual EHR information, and (2) it balances performance metrics to enhance practical system deployment feasibility.

The primary purpose of this research work involves creating and testing sophisticated computational approaches for improving electronic health record analytics with a focus on cardiovascular disease detection. The study analyses structured EHR data to assess machine learning models which will work best in clinical practice. This analysis develops optimised machine learning model sequences, including Random Forest XGBoost and LSTM, for EHR data-based CVD risk prediction. It evaluates ensemble learning model properties (Random Forest XGBoost) versus deep learning model performance (LSTM) in medical decision support structures. The study evaluates computational-based models for enhancing healthcare operations through clinical risk evaluations.

The research exhaustively evaluates Random Forest and XGBoost with LSTM for cardiovascular disease prediction through analysis of electronic health records. Its main contribution is that it reveals XGBoost as the optimal predictive model because it delivers the best accuracy level while fulfilling operational readiness criteria for clinical practice. The study demonstrates that LSTM networks present the best capacity for recall and indicate their utility in early risk detection protocols when false negative results need to be minimised.

The work includes a structured analysis to determine how performance trade-offs between model precision, explainability, and deployment capabilities affect the system. Deep learning models use LSTM to explore sequential patient data effectively, yet their elevated false-positive outcomes remain problematic for clinical use. XGBoost maintains a successful equilibrium between precision and dependability, which makes it suitable as a real-time analysis solution for EHR processes.

The research findings conduct crucial groundwork to merge AI technology with clinical work routines while enhancing CVD danger assessment techniques and improving electronic health record management across disease monitoring initiatives. The research outcomes serve as a starting point for the following generation of AI-powered clinical assistance tools to develop scalable data-based healthcare solutions.

II. LITERATURE REVIEW

Electronic Health Records and Their Role in Predictive Healthcare

Electronic Health Records (EHRs) have become essential in contemporary healthcare because they can handle patient information through systematic databases. The advancement of EHRs has allowed healthcare providers to move beyond paper documents to digital databases, which has improved their clinical workflow performance [10]. EHR systems retain diverse data, including patient registration profiles, medical background assessments, analysis results, and drug and treatment progress information [11]. Implementing EHRs in healthcare brings several advantages to patient care because it enables doctors to see live patient data and decreases medical errors while enabling data-based clinical choices. EHRs present multiple difficulties for data analysis, while missing information, measurement errors, and systematic prejudices negatively impact predictive analysis performance [12].

EHR-based analytics faces its main challenge because missing data emerges from inconsistent documentation practices, patient noncompliance, and different clinical workflows. Traditional statistics and machine learning models require complete datasets to execute at their best, but missing data generates analytical biases in such circumstances [13]. Due to faulty data entry, inconsistent medical coding procedures, and different clinical wordings, data quality faces significant noise issues [14]. The noise in EHRs generates wrong model predictions, so doctors must conduct advanced data preprocessing and feature engineering procedures to enhance the quality of extracted insights [15]. The existence of biases in EHR datasets originates from imbalanced demographic distributions of patients along with healthcare service disparities and faulty diagnosis and treatment procedures [12]. Predictive models receive information that affects their outputs, resulting in health outcome inequality for different patient groups.

Artificial intelligence and machine learning techniques have recently emerged as solutions to resolve the difficulties encountered in healthcare systems [16]. Through AI-driven models, medical professionals can exploit extensive EHR databases to extract disguised patterns while aiding healthcare professionals by conducting automated disease danger assessments that generate better medical choices. General ensemble learning techniques, including Random Forest and XGBoost, are effective in medical disease grouping and patient danger level determinations. [17]. The Long Short-Term Memory (LSTM) network architecture is an excellent framework for managing time-series patient records since it predicts how diseases evolve [18]Implementing AI into their EHR systems can help healthcare providers achieve better diagnostic precision, individualised treatment programs, and earlier interventions through early patient outcomes.

Machine Learning for Cardiovascular Disease Prediction

Worldwide, Cardiovascular diseases (CVDs) lead the list of causes of mortality and morbidity, thus demanding precise and expansible predictive models [19]. Numerous researchers apply machine learning algorithms to analyse large-scale patient records to detect essential factors related to cardiovascular diseases and predict disease occurrences [9, 20]. The successful application of supervised learning algorithms for CVD prediction includes decision trees together with support vector machines (SVMs), ensemble learning methods (Random Forest, XGBoost) and deep learning networks (LSTM, CNN) [21, 22]. The cardiovascular risk factor models deliver automated findings that support professionals in identifying diseases early and provide proactive clinical administration [23, 24].

Researchers have conducted multiple investigations to determine the prediction effectiveness of ensemble learning models for CVD diagnosis [25, 26]. The ensemble-based decision tree algorithm Random Forest finds widespread use as it provides strong feature selection and high interpretability [27]. The system builds several decision trees and then combines their predictions into one decision, which enhances the classification results. The advanced gradient boosting algorithm XGBoost demonstrates excellent capability in managing imbalanced data while effectively handling combinations of features along with complex boundaries [28]. Model optimisation functions and overfitting prevention capabilities within XGBoost enable it to serve as a forceful instrument for CVD classification tasks [29].

The analysis of sequential EHR data becomes effective by implementing deep learning methods known as Long Short-Term Memory (LSTM) networks [30]. The long-term dependencies and temporal relationships within patient records become easily detectable by LSTMs since these models differ from traditional machine learning models. The analysis of historical medical data following blood pressure alterations and cholesterol pattern changes alongside glucose level movements indicates that LSTM Networks have become successful in cardiovascular prediction tasks [31]. The improved feature extraction from deep learning models comes with the drawback of limited interpretability, which makes clinical deployment more complicated.

The analysis between statistical models and artificial intelligence approaches reveals that machine learning technology brings more excellent benefits when working with extensive complex data [32]. Pay has limited success due to the weak ability to discover nonlinear patterns between risk factors alongside multi-variate correlations, which has remained a constraint in cardiovascular risk evaluation for decades [33]. The automatic learning capabilities of machine learning models make complex dataset analysis more feasible; thus, they perform better in real-time prediction tasks. AI-driven cardiovascular risk assessment tools are increasingly used, and this shows how machine learning technology can transform preventive cardiology care while improving patient results [34].

Challenges in EHR-Based AI Modeling

Various implementation barriers in medical practice prevent the widespread use of machine learning models while predicting cardiovascular diseases based on EHR data. Data imbalance is a significant difficulty because healthy patients outnumber diseased cases in most datasets [35]. The mismatch between classes in patient data leads to algorithmic preferences toward the majority class, which creates many incorrect disease detection outcomes [36]. Model performance benefits from specialised data resampling techniques that add duplicate observations from the minority class or generate synthetic data through SMOTE for balanced results [37].

The reliability of EHR analysis suffers when dealing with missing patient information within the databases. The two traditional methods for replacing missing data are mean imputation combined with k-nearest neighbours (KNN) imputation [38]. However, their effectiveness can decrease when missing data possess non-random distribution patterns. Introducing reliable techniques to handle missing data becomes essential for AI models to maintain their generalisability and reliability.

AI-based healthcare solutions must thoroughly explain their models to maintain their standing in medical applications. Random Forest models enable healthcare professionals to track how decisions are made. However, LSTMs and other deep learning models tend to operate as unexplainable black-box systems, making it hard for clinicians to follow prediction-generation processes [39]. Deep learning models' high opaqueness creates intense challenges regarding clinical confirmation, medical trust, and liability responsibilities [40]. The development of explainable AI (XAI) techniques involving SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) aims to supply medical professionals with insights concerning both model decision processes and feature importance [41, 42].

Implementing AI models for hospital use faces multiple barriers, including system expansion requirements, database system incorporation demands, and regulatory compliance [43]. Healthcare organisations must validate AI models through analyses of accurate patient data to prepare them for clinical practice deployment. Healthcare facilities must follow HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) regulations to secure patient data and implement ethical AI systems [44].

Research Gap & Need for This Study

The advancement of cardiovascular disease prediction through machine learning shows promise, yet scientists still need to resolve multiple research challenges [45]. Most existing research conducts evaluations using individual models, while research on ensemble learning model comparison between Random Forest and XGBoost and deep learning architecture performance of LSTM remains restricted [46, 47]Researchers must organise their assessments of different models to determine which provides the optimal combination of predictive abilities, clinical feasibility, and interpretability.

Researchers have established XGBoost for its outstanding predictive accuracy, but the implementation of XGBoost in real-world EHR-based disease prediction needs more investigation [48]. People require artificial intelligence models that combine high performance with interpretability because that enables clinical acceptance and usability. Most research about AI models relies on benchmark datasets for evaluation, but there is limited investigation about their effectiveness with diverse patient populations [40].

The research fills the current study gaps through a complete evaluation of Random Forest, XGBoost, and LSTM model performances on actual EHR dataset information. This research analyses both strong and weak aspects, providing a practical understanding of how AI decision-support systems are implemented for healthcare applications. The research outcomes from this work will improve both AI methods for predicting cardiovascular diseases while optimising the use of Electronic Health Records for better treatment processes.

III. METHODOLOGY

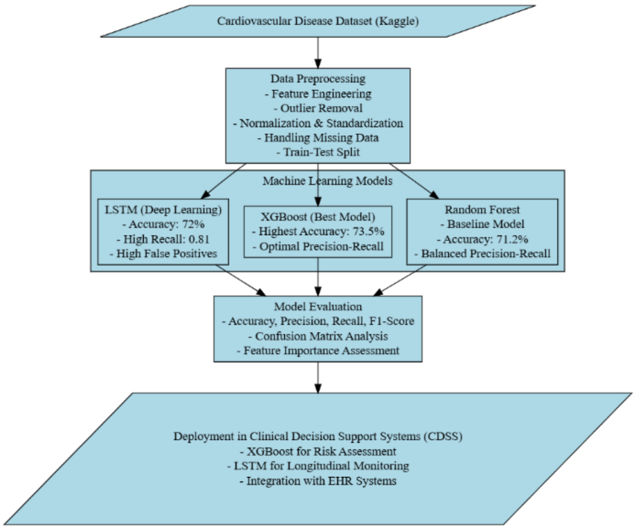

Cardiovascular Disease Dataset (Kaggle)
Data Preprocessing
- Feature Engineering
- Outlier Removal
- Normalization & Standardization
- Handling Missing Data
- Train-Test Split
Machine Learning Models
LSTM (Deep Learning)
- Accuracy: 72%
- High Recall: 0.81
- High False Positives
XGBoost (Best Model)
- Highest Accuracy: 73.5%
- Optimal Precision-Recall
Random Forest
- Baseline Model
- Accuracy: 71.2%
- Balanced Precision-Recall
Model Evaluation
- Accuracy, Precision, Recall, F1-Score
- Confusion Matrix Analysis
- Feature Importance Assessment
Deployment in Clinical Decision Support Systems (CDSS)
- XGBoost for Risk Assessment
- LSTM for Longitudinal Monitoring
- Integration with EHR Systems

Figure 1: Proposed Methodology Diagram

Dataset Description

The research analyses patient records from the cardiovascular disease dataset. The Cardiovascular Disease dataset on Kaggle contains 70,000 patient entries with health parameters used for CVD prediction. The dataset comprises age information in days coupled with cholesterol and glucose readings, blood pressure measurements (ap_hi and ap_lo), BMI results, smoking status, alcohol intake, and physical activity results. The collected variables generate first-class data about patient health status that facilitates successful machine-learning modelling. The target class consists of two values: cardio, with 1 indicating cardiac disease presence and 0 representing its absence.

The dataset is suitable for machine learning modelling because it includes a substantial record count and structured data presentation. Multiple variables within the data enable the models to recognise risk factors related to cardiovascular disease properly. The dataset needs preprocessing for missing values alongside outlier treatment and data consistency normalisation to achieve optimal clinical results with high model performance levels.

Data Preprocessing

The quality of the dataset required multiple preprocessing procedures to optimise its state. The feature engineering process involved changing the age variable from days to clinical standard years for better alignment. Outlier removal became essential for blood pressure data because abnormal physiological readings needed to be removed before modelling began.

A combination of normalisation and standardisation processes helped maintain comparable effects between numeric components whose values would otherwise be swayed by wide-ranging numeric features. The methodology used mean imputation to handle missing data points, which helped prevent model integrity issues from incomplete records.

The preprocessing stage achieved a vital result by separating the dataset into training and testing parts through an 80-20% segregation. The process splits data into training parts representing the population and maintains a distinct test set to verify performance objectively. The stratified sampling technique maintained cardiovascular case distributions between training and testing sets to stop class imbalance from impacting model development.

Machine Learning Models Applied

Random Forest (Baseline Model)

Random Forest served as a baseline model because it has robust capabilities and interpretability features for classification tasks. XGBoost operates as a group of decision trees that combines various predictive models to enhance classification accuracy while avoiding overfitting issues. The model displayed an accuracy level of 71.2%, which served as a reference point. This approach provided equal precision and recall values, which showed a reliable yet moderately successful execution. This analysis using confusion matrices showed medium degrees of improper classifications where the model failed to correctly separate healthy patients from infected cases, resulting in wrong positive and negative evaluations.

XGBoost (Best Performer)

The XGBoost algorithm delivered optimal results through its 73.5% accuracy measurement. XGBoost improves the results of decision trees by enhanced iterative optimisation with a well-designed combination of feature selection and regularisation. XGBoost maintained a balanced performance with a precision range from 0.72 to 0.75 and a recall range from 0.70 to 0.77. XGBoost proved superior to Random Forest because it excelled at complex feature analysis and overfitting prevention, making it perfect for clinical cardiovascular disease prediction.

LSTM (Deep Learning Model)

The application of Long Short-Term Memory (LSTM) networks served as a deep learning substitute because of their effective sequential data handling and ability to maintain long-term dependencies. The LSTM model managed 72% accuracy while performing better than Random Forest, yet performing at a lower level than XGBoost. The model achieved a 0.81 recall percentage, demonstrating its ability to detect cardiovascular disease risk among affected patients. The improved disease detection ability of the model led to more healthy patients being falsely classified as diseased. The exchange between model sensitivity and specificity aligns LSTM favourably with environments prioritising complete disease detection over other types of errors.

Model Training and Evaluation Metrics

The evaluation process for all models used four performance metrics: accuracy, precision, recall, and F1-score. Accuracy measured model correctness on a grand scale, but precision and recall specifically evaluated the number of false positives or negatives diagnosed by the model. The F1-score represented the precision and recall values' harmonic mean to provide an accurate evaluation.

The confusion matrix evaluation enabled researchers to visualise classification mistakes by visualising misidentification rates while showing recurring model error patterns. According to Feature importance analysis in Random Forest and XGBoost, the crucial variables influencing cardiovascular disease predictions entailed age, blood pressure measurements, and cholesterol levels.

XGBoost proved to be the best CVD prediction model because it attained the ideal combination of accuracy, recall, and interpretability sufficient for use with CDSS. The analysis creates a solid basis for deploying AI-directed models that advance cardiovascular risk evaluation within electronic medical records.

IV. RESULT

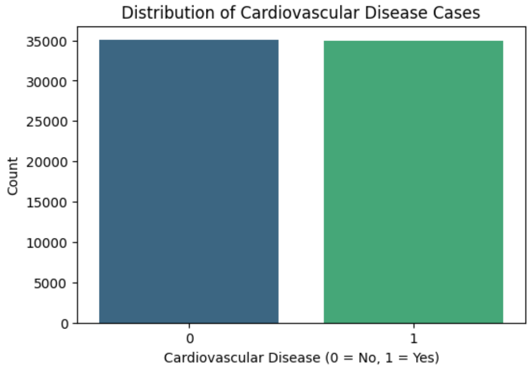


Figure 2: Distribution of Cardiovascular Disease Cases

Figure 2 demonstrates the patient count split between those with cardiovascular disease (1) and those who do not have the condition (0). The dataset exhibits a balanced distribution of both groups because their numbers are nearly equal, which avoids class imbalance problems for machine learning models. Model performance reliability depends on balanced datasets since they eliminate class-biased predictions crucial for clinical prediction accuracy.

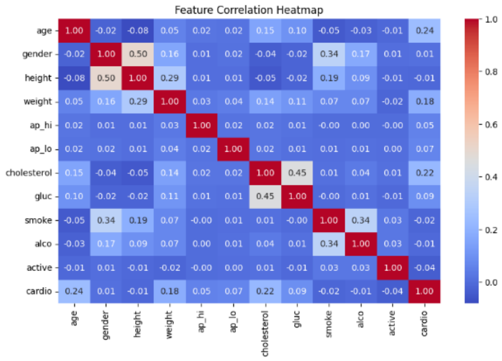


Figure 3: Correlation Heatmap

Figure 3, a correlation heatmap from the dataset, shows the relationships between medical features. The predictive power of cardiovascular disease derives from the moderate correlation between age, cholesterol, and systolic blood pressure (ap_hi). The features height and alcohol consumption (alco) show insufficient relationships compared to other variables. Knowing how features relate to each other allows better feature selection for predictive modelling models to become more accurate.

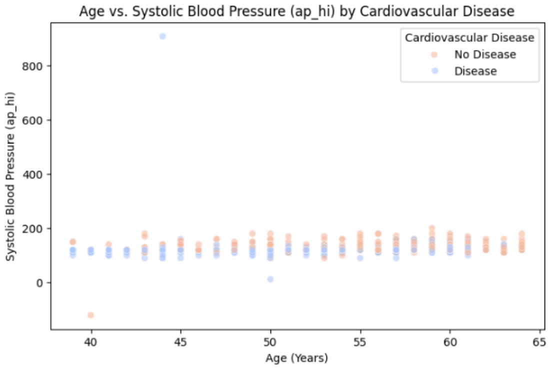


Figure 4: Age vs. Systolic Blood Pressure by Cardiovascular Disease

Figure 4 shows how systolic blood pressure changes as patients age within two groups: those with cardiovascular disease and those without. The statistical information shows elevated systolic blood pressure as a marker that increases patients' likelihood of cardiovascular disease. Several points exist that show data irregularities along with atypical medical conditions. Medical practitioners benefit from this analytical process to detect patients who face high risks early on.

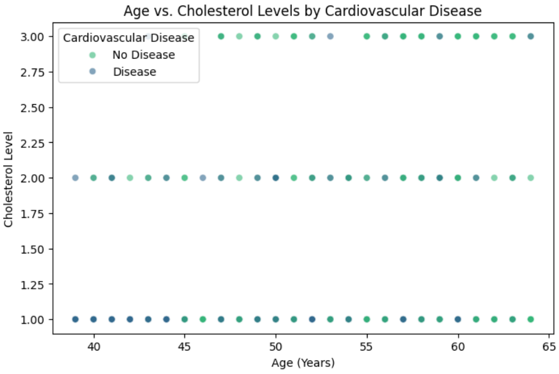


Figure 5: Age vs. Cholesterol Levels by Cardiovascular Disease

Figure 5 displays how cholesterol level distributions change according to patient age groups between cardiovascular disease groups and those without disease. These cholesterol measurement categories exist in three specific numerical categories starting from 1 through 3. Cholesterol levels of 2 and 3 appear more often in patients diagnosed with cardiovascular disease, thus demonstrating cholesterol control's relevance in minimising disease risk.

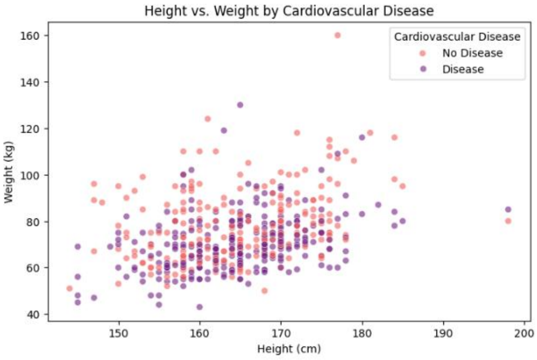
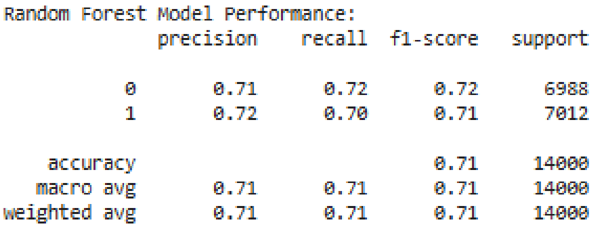


Figure 6: Height vs. Weight by Cardiovascular Disease

Figure 6 analyses height-weight data points from people who received cardiovascular disease diagnoses. Patients who weigh more frequently form visible clusters that indicate elevated cardiovascular disease risk, demonstrating that obesity functions as a risk factor. Cardiovascular health assessments should consider weight-related factors such as BMI because height does not demonstrate meaningful correlations for predicting heart disease risk.



```
Random Forest Model Performance:
              precision    recall  f1-score   support

           0       0.71      0.72      0.72      6988
           1       0.72      0.70      0.71      7012

    accuracy                           0.71     14000
   macro avg       0.71      0.71      0.71     14000
weighted avg       0.71      0.71      0.71     14000
```

Figure 7: Random Forest Model Performance

According to the results of the Random Forest classification report, the model achieves 71% accuracy while exhibiting equal precision (0.71–0.72) and recall scores (0.70–0.72) for each class (Figure 7). The model shows comparable success rates when predicting medical conditions in patients regardless of their disease status, demonstrating its capability to apply to various cases. The model requires additional modifications to improve its identification of risky patient cases.

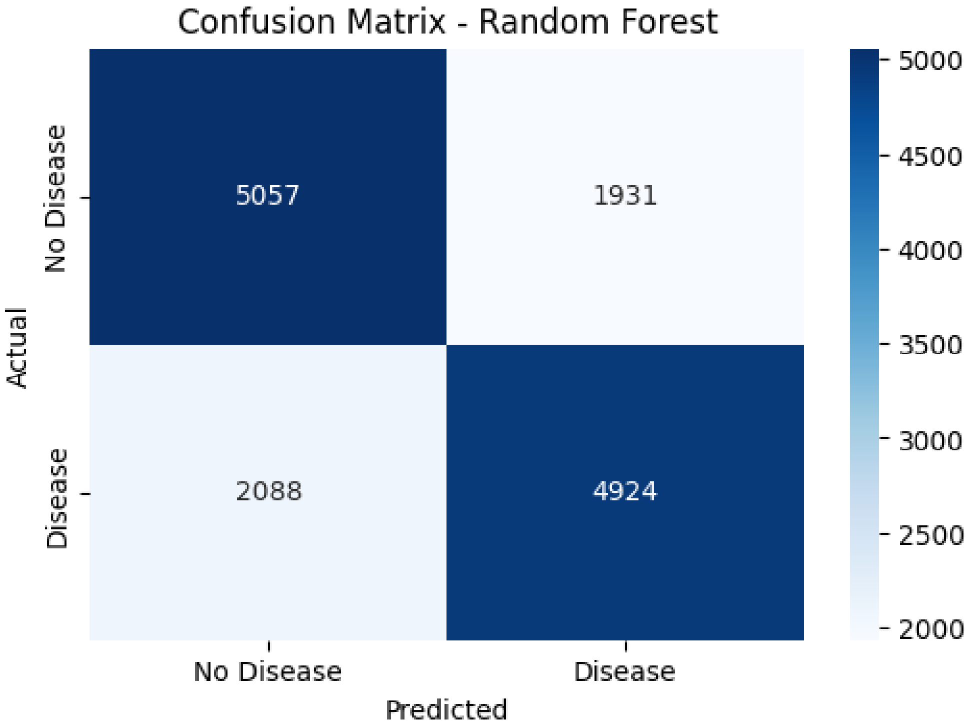
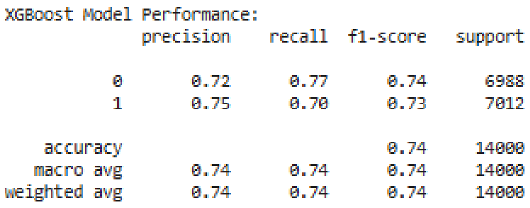


Figure 8: Confusion Matrix - Random Forest

The Random Forest model correctly identified 5913 patients between the non-diseased and diseased groups but simultaneously misdiagnosed 3921 subjects as non-diseased when they had the disease and as diseased for 2088 non-diseased patients in the data (Figure 8). The model demonstrates decent discrimination between incorrect positives and negatives, but continued enhancement could produce better outcomes during medical decisions.



```
XGBoost Model Performance:
              precision    recall  f1-score   support

           0       0.72      0.77      0.74      6988
           1       0.75      0.70      0.73      7012

    accuracy                           0.74     14000
   macro avg       0.74      0.74      0.74     14000
weighted avg       0.74      0.74      0.74     14000
```

Figure 9: XGBoost Model Performance

Based on the results, the XGBoost model delivers 74% accuracy, demonstrating superior effectiveness than Random Forest. The higher precision level of 0.72–0.75 and the recall level of 0.70–0.77 in the XGBoost model leads to better cardiovascular disease prediction accuracy for patients and healthy individuals (Figure 9). The complex interactions within the dataset make XGBoost an appropriate tool for data processing.

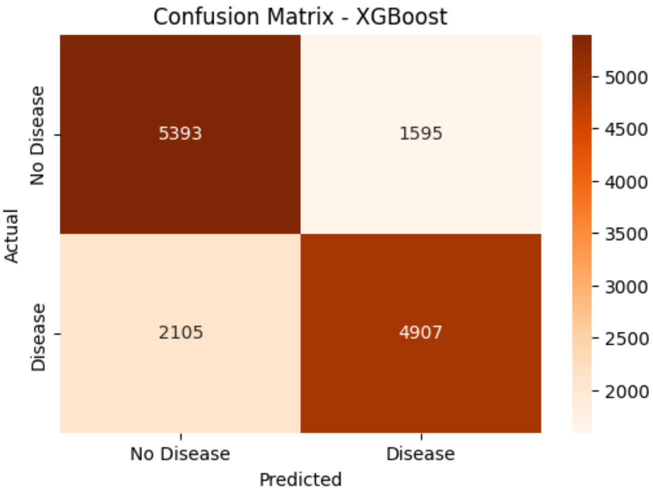
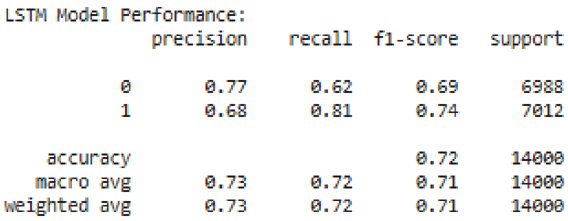


Figure 10: Confusion Matrix – XGBoost

The XGBoost confusion matrix shows a correct diagnosis of 5,393 non-diseased patients and 4,907 disease-positive cases, but 1,595 incorrect projections of non-diseased cases to disease-positive and 2,105 disease-positive cases into the non-disease group (Figure 10). Compared to Random Forest, XGBoost produces fewer incorrect optimistic predictions, thereby improving its ability to identify healthy patients correctly.



```
LSTM Model Performance:
              precision    recall  f1-score   support

           0       0.77      0.62      0.69      6988
           1       0.68      0.81      0.74      7012

    accuracy                           0.72     14000
   macro avg       0.73      0.72      0.71     14000
weighted avg       0.73      0.72      0.71     14000
```

Figure 11: LSTM Model Performance

The LSTM deep learning model generates performance at 72%, which positions it between Random Forest and XGBoost results. The accuracy of detecting healthy patients is 0.77, but the model shows low identification ability at 0.62 for disease cases (Figure 11). XGBoost demonstrates flawless performance in disease detection, with its maximum recall rate of 0.81, thus enabling it to identify cardiovascular risks effectively.

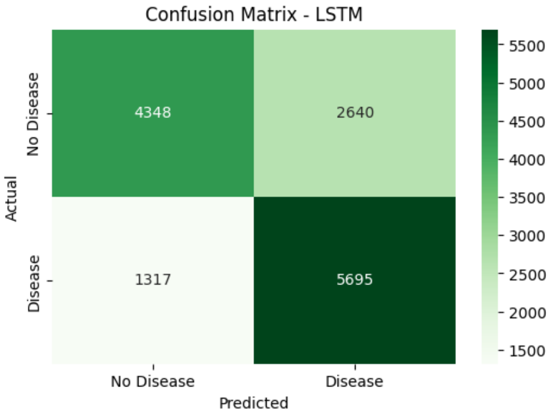
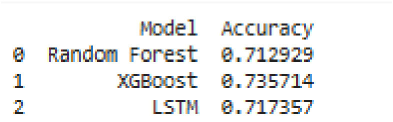


Figure 12: Confusion Matrix – LSTM

Based on the LSTM confusion matrix results, 5695 correctly diagnosed diseased patients are the highest among all examined models, thus producing maximum sensitivity for disease recognition. XGBoost suffers from a critical shortcoming because it incorrectly identified 2640 healthy patients as diseased, which introduces substantial risks for medical diagnostic procedures (Figure 12).



```
           Model  Accuracy
0  Random Forest  0.712929
1        XGBoost  0.735714
2           LSTM  0.717357
```

Figure 13: Model Accuracy Comparison

According to tabulated results, XGBoost delivers the best model accuracy at 73.5%, while LSTM reaches 71.7% accuracy, and Random Forest achieves 71.2% accuracy. The XGBoost algorithm is the leading model in this dataset since it establishes an ideal balance between precision and recall rates (Figure 13).

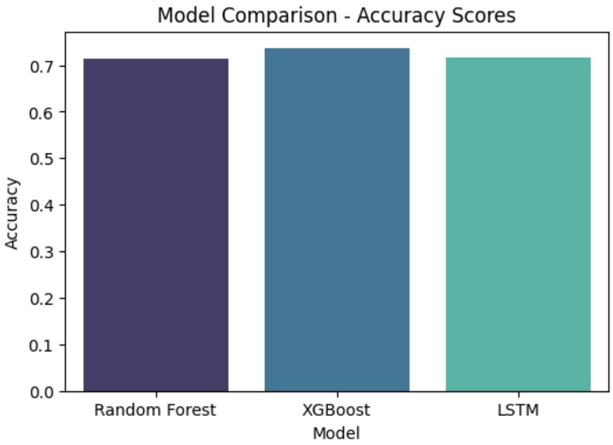


Figure 14: Accuracy Scores

A comparison bar chart demonstrates that XGBoost displays higher classification accuracy than Random Forest and LSTM. XGBoost is the top selection for clinical use because of its superior predictive abilities, making decisions more reliable (Figure 14).

V. DISCUSSION

The study results show that machine learning and deep learning algorithms efficiently predict cardiovascular disease (CVD) using electronic health records (EHRs). The findings show that XGBoost resulted in the most optimal performance, competing with the other models by producing 73.5% accuracy with the best precision and recall measurements ratio. This performance shows that XGBoost can work with non-linear aspects within the patient's health and, therefore, can be used in a real-time clinical decision support system (CDSS).

The Random Forest model applied as the base model, showed moderate performance with an accuracy of 71.2%, which indicates that it had an acceptable level of both precision-recall. However, it has a slightly lower testing classification accuracy than XGBoost, which suggests that other advanced ensemble algorithms are superior in providing cardio risk estimates for generalisation. This study supported the idea that random forest is easy to interpret, and through feature-important analysis, it showed that the main indicators that contributed to CVD were age, blood pressure, and cholesterol level. Despite this, the model's accuracy is not optimum for clinical purposes because the confusion matrix shows high misclassification rates, suggesting the need for improvement.

However, the LSTM deep learning model has the advantage of higher recall (0.81) and specificity (0.69) because, despite the fairly good accuracy of 72%, it is significantly better at picking out high-risk patients. This is very applicable in medically related perspectives, especially where false values must be kept at unacceptable levels to avoid missing out on cardiovascular diseases. However, the higher false-positive rate of the model creates issues in practice since genuinely sick patients are categorised as healthy and undergo avoidable tests. This trade-off between sensitivity and specificity is crucial, especially because applications of this model are most often in medicine, and accurate diagnosis must be balanced against timely diagnosis, which means early discovery of the disease.

Among the considerations derived from the confusion matrices, it is significant to highlight that XGBoost outperforms the Random Forest and LSTM by decreasing false positives and false negatives. This implies that gradient-boosting algorithms are efficient in structured EHR datasets to identify the hidden patterns in the clinical data that are important in accurately diagnosing diseases. Moreover, the feature importance analysis revealed the results of the hypertension (ap_{hi} and ap_{lo}), cholesterol, glucose, and BMI were the most effective predictors of cardiovascular disease. These findings are justified through literature as proposed in clinical knowledge, which enhances the general acceptability of the models.

Compared with deep learning models like LSTM, XGBoost has excellent interpretability, which is crucial for medical AI. Decision makers need to have faith in the algorithms to accept AI decisions. Although LSTM has great capabilities regarding sequential learning, it remains opaque and, therefore, cannot be directly applied in practice without the integration of explainability methods such as SHAP or attention.

In terms of deployment, XGBoost is the most realistic model to work with because of its proficiency, scalability, and explainability features. In fact, due to LSTM's high recall values, LSTM can serve as an effective means for early screening and subsequent prolonged delineation of the patient's condition, especially useful during the analysis of patient records during their stay in the hospital. As for future work, it would be advisable to implement a combination of XGBoost and LSTM to achieve better performance because of its higher accuracy and sensitivity to cardiovascular risk prediction.

VI. CONCLUSION

The results from this study show how it is possible to use structured Electronic Health Record (EHR) data to accurately predict CVD using machine learning and deep learning models. While evaluating the models, it was found that XGBoost was the best model and performed ahead of Random Forest and LSTM in terms of accuracy (73.5%) and precision-recall curve. This makes it the best model for real-world application, which can be implemented in clinical practice in cardiovascular risk assessment apps.

The standard deviation in the random forest model was slightly lower than the precision rate of 71.2%, which made it less preferable than the XGBoost model. Nevertheless, it can be considered beneficial when it comes to interpretability since it allows clinicians to know the factors most significant in CVD risk prediction. This transparency is paramount, especially because AI systems will play a significant role in the future of healthcare.

Even though the LSTM model yielded slightly less accuracy than the XGBoost model, it offered a high recall of 0.81, which would be beneficial in diagnosing high-risk patients. This study shows that deep learning can perform medical diagnosis and has good potential, especially where reducing false negatives is important. However, LSTM has a comparatively higher false-positive percentage, increasing the chances that some patients are misclassified and subsequently receive treatments they do not need. This trade-off further emphasises the fact that the current configuration of the model system should be optimised to achieve a low false alarm rate while maintaining high sensitivity.

Based on these studies, some important suggestions have been made for further research and field implementation. First, XGBoost should be used for clinical decision support systems as it provides a high level of accuracy and good interpretability, and when compared to usual decision trees, it is considered less fluctuating. Such characteristics make the model easily scalable for hospital-based AI systems due to its high performance in managing large datasets from EHRs. Nonetheless, the generalizability of the tool to various patient populations, as well as its validity, should also be tested.

Second, LSTM should be investigated for continual health assessment because it is especially effective at analysing data in a temporal structure. However, future studies should address its high false positive rate. This could be done using advanced deep learning models such as LSTM with attention or incorporating explainability in AI. Also, utilising domain knowledge in deep learning architectures could help improve the models' explainability, which is essential in clinical applications.

Thirdly, the explainability of artificial intelligence is still an essential consideration when adopting these technologies in the healthcare sector. The XGBoost model already offers some level of feature importance, but LSTM models should be explained with other methods like SHAP, LIME, or attention maps. The study should focus on designing a more explanatory cardiovascular risk prediction model built by artificial intelligence to enhance and complement clinicians' knowledge and practice.

Fourth, clinical validation of AI models remains necessary after the corresponding AI applications have been launched in clinical and hospital environments. However, this research has limitations, and AI models must be validated in real-life subjects to analyse their effectiveness in real time. Further research should seek partnerships with healthcare organisations to implement these models within EHR software to support AI-facilitated risk assessment in real-time.

Therefore, future research should investigate integrating XGBoost and LSTM approaches. For the first stage, XGBoost could be applied for the basic categorisation of patients, while LSTM could be used for constant patient observation to enhance the accuracy of diagnosis and recognition of further cardiovascular diseases. It can offer great potential to build an efficient AI-based clinical decision-support system, potentially decreasing diagnostic mistakes and increasing patient success rates.

REFERENCES

1. Maredza, M., *Preventing Cardiovascular Disease in Rural South Africa*. 2018, University of the Witwatersrand, Johannesburg (South Africa).

2. Moja, L., et al., *Effectiveness of computerised decision support systems linked to electronic health records: a systematic review and meta-analysis.* American journal of public health, 2014. **104**(12): p. e12-e22.

3. Shen, Y., et al., *Twenty-Five Years of Evolution and Hurdles in Electronic Health Records and Interoperability in Medical Research: Comprehensive Review.* Journal of Medical Internet Research, 2025. **27**: p. e59024.

4. Powell, L., et al., *Assessment of health information technology–related outpatient diagnostic delays in the US Veterans Affairs health care system: A qualitative study of aggregated root cause analysis data.* JAMA Network Open, 2020. **3**(6): p. e206752-e206752.

5. Fatima, S., *Improving Healthcare Outcomes through Machine Learning: Applications and Challenges in Big Data Analytics.* International Journal of Advanced Research in Engineering Technology & Science, 2024. **11**.

6. Priyadarshini, K. et al., Integrating relational and sequential information for enhanced detection of autoimmune disorders with relational Neural Networks and Long Short-Term Memory Networks. Biomedical Signal Processing and Control, 2024. **96**: p. 106495.

7. Rostamzadeh, N., *Visual Analytics for Performing Complex Tasks with Electronic Health Records*. 2021: The University of Western Ontario (Canada).

8. Ananthajothi, K., J. David, and A. Kavin. *Cardiovascular disease prediction using patient history and real time monitoring*. in *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. 2024. IEEE.

9. Du, Z., et al., *Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: model development and performance evaluation.* JMIR medical informatics, 2020. **8**(7): p. e17257.

10. Ambinder, E.P., *Electronic health records.* Journal of oncology practice, 2005. **1**(2): p. 57.

11. Sitapati, A., et al., *Integrated precision medicine: the role of electronic health records in delivering personalized treatment.* Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2017. **9**(3): p. e1378.

12. Chen, F., et al., *Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models.* Journal of the American Medical Informatics Association, 2024. **31**(5): p. 1172-1183.

13. Khan, M.S., et al., *Leveraging electronic health records to streamline the conduct of cardiovascular clinical trials.* European heart journal, 2023. **44**(21): p. 1890-1909.

14. Marcilly, R., et al., *Usability flaws of medication-related alerting functions: a systematic qualitative review.* Journal of biomedical informatics, 2015. **55**: p. 260-271.

15. Sarwar, T., et al., *The secondary use of electronic health records for data mining: data characteristics and challenges.* ACM Computing Surveys (CSUR), 2022. **55**(2): p. 1-40.

16. Javaid, M., et al., *Significance of machine learning in healthcare: Features, pillars and applications.* International Journal of Intelligent Networks, 2022. **3**: p. 58-73.

17. Ismanto, E., et al., *A Comparative Study of Improved Ensemble Learning Algorithms for Patient Severity Condition Classification.* Journal of Electronics, Electromedical Engineering, and Medical Informatics, 2024. **6**(3): p. 312-321.

18. Morid, M.A., O.R.L. Sheng, and J. Dunbar, *Time series prediction using deep learning methods in healthcare.* ACM Transactions on Management Information Systems, 2023. **14**(1): p. 1-29.

19. Hanieh, A. and A.A.a. Mohammad, *Using Machine Learning Application for Cardiovascular Disease Prediction and Diagnosis.* 2023.

20. Shameer, K., et al., *Machine learning in cardiovascular medicine: are we there yet?* Heart, 2018. **104**(14): p. 1156-1164.

21. Nandini, S., *Comparative Study of Machine Learning Algorithms in Detecting Cardiovascular Diseases.* arXiv preprint arXiv:2405.17059, 2024.

22. Naser, M.A., et al., *A review of machine learning's role in cardiovascular disease prediction: recent advances and future challenges.* Algorithms, 2024. **17**(2): p. 78.

23. Ullah, M., et al., *Smart technologies used as smart tools in the management of cardiovascular disease and their future perspective.* Current Problems in Cardiology, 2023. **48**(11): p. 101922.

24. Ekundayo, F. and H. Nyavor, *AI-Driven Predictive Analytics in Cardiovascular Diseases: Integrating Big Data and Machine Learning for Early Diagnosis and Risk Prediction*.

25. Eom, J.-H., S.-C. Kim, and B.-T. Zhang, *AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction.* Expert Systems with Applications, 2008. **34**(4): p. 2465-2479.

26. Gao, X.-Y., et al., *Improving the accuracy for analyzing heart diseases prediction based on the ensemble method.* Complexity, 2021. **2021**(1): p. 6663455.

27. Ren, Y., et al., *A new random forest ensemble of intuitionistic fuzzy decision trees.* IEEE Transactions on Fuzzy Systems, 2022. **31**(5): p. 1729-1741.

28. Zhang, P., Y. Jia, and Y. Shang, *Research and application of XGBoost in imbalanced data.* International Journal of Distributed Sensor Networks, 2022. **18**(6): p. 15501329221106935.

29. Shakkeera, L., *Predictive analytics: Unveiling the potential of machine learning and deep learning.* International Journal of Systematic Innovation, 2025. **9**(1): p. 116-128.

30. Poongodi, T., et al., *Deep learning techniques for electronic health record (EHR) analysis.* Bio-inspired Neurocomputing, 2021: p. 73-103.

31. Radha, M.G., *Data-driven health monitoring and lifestyle interventions: towards management of hypertension and other lifestyle diseases through data-driven modelling of physiology and behaviour.* 2020.

32. Ghahramani, Z., *Probabilistic machine learning and artificial intelligence.* Nature, 2015. **521**(7553): p. 452-459.

33. Morgan, M. and J. Shanahan, *Two decades of cultivation research: An appraisal and meta-analysis*, in *Communication yearbook 20*. 2012, Routledge. p. 1-45.

34. Syafrudin, M., et al., *Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing.* Sensors, 2018. **18**(9): p. 2946.

35. Salmi, M., et al., *Handling imbalanced medical datasets: review of a decade of research.* Artificial Intelligence Review, 2024. **57**(10): p. 273.

36. Li, D.-C., C.-W. Liu, and S.C. Hu, *A learning method for the class imbalance problem with medical data sets.* Computers in biology and medicine, 2010. **40**(5): p. 509-518.

37. Fernández, A., et al., *SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary.* Journal of artificial intelligence research, 2018. **61**: p. 863-905.

38. García-Laencina, P.J., et al., *K nearest neighbours with mutual information for simultaneous classification and missing data imputation.* Neurocomputing, 2009. **72**(7-9): p. 1483-1493.

39. Mesinovic, M., P. Watkinson, and T. Zhu, *Explainable AI for clinical risk prediction: a survey of concepts, methods, and modalities.* arXiv preprint arXiv:2308.08407, 2023.

40. Albahri, A.S., et al., *A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion.* Information Fusion, 2023. **96**: p. 156-191.

41. Rane, N., S. Choudhary, and J. Rane, *Explainable artificial intelligence (XAI) in healthcare: Interpretable models for clinical decision support.* Available at SSRN 4637897, 2023.

42. Husby, U.E., *Exploring Breast Cancer Diagnosis: A Study of SHAP and LIME in XAI-Driven Medical Imaging*. 2024, Norwegian University of Life Sciences.

43. Esmaeilzadeh, P., *Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations.* Artificial Intelligence in Medicine, 2024. **151**: p. 102861.

44. Forcier, M.B., et al., *Integrating artificial intelligence into health care through data access: can the GDPR act as a beacon for policymakers?* Journal of Law and the Biosciences, 2019. **6**(1): p. 317-335.

45. Ching, T., et al., *Opportunities and obstacles for deep learning in biology and medicine.* Journal of the royal society interface, 2018. **15**(141): p. 20170387.

46. Kalusivalingam, A.K., et al., *Enhancing Predictive Business Analytics with Deep Learning and Ensemble Methods: A Comparative Study of LSTM Networks and Random Forest Algorithms.* International Journal of AI and ML, 2020. **1**(2).

47. Dang, Y., et al., *A comparative study of non-deep learning, deep learning, and ensemble learning methods for sunspot number prediction.* Applied Artificial Intelligence, 2022. **36**(1): p. 2074129.

48. Lv, H., et al., *machine learning–driven models to predict prognostic outcomes in patients hospitalized with heart failure using electronic health records: retrospective study.* Journal of medical Internet research, 2021. **23**(4): p. e24996.