Heart Disease Prediction Using Machine Learning

Nikkitha G[1], Nivedha N[2], Raama Nivethaa D[3], Sree Harshini S[4], Mrs.R Divya[5]

[1, 2, 3, 4]UG Scholars, [5]Assistant Professor, Department of Computer Science and Engineering,

Avinashilingam Institute for Home Science and Higher Education for Women, School of Engineering, Coimbatore, India

ABSTRACT

Heart failure is a chronic disease affecting millions worldwide. An efficient machine learning- based technique is needed to predict heart failure health status early and take necessary actions to overcome this worldwide issue. The study addresses data preparation difficulties such as missing values and outliers by leveraging a large dataset including critical health parameters such as age, gender, blood pressure, and cholesterol levels. The models are trained using this new dataset, and the anticipated performance is rigorously assessed using standard criteria. The study includes a detailed review of the Warm and Naive Bayes (NB) models, highlighting their benefits and drawbacks in the context of heart disease prediction. Our proposed research study has significant scientific contributions to the medical community.

Keywords: Heart Disease, Machine Learning, Predictive Modeling, Medical History

1.INTRODUCTION

Heart disease is a significant worldwide cause of mortality, posing a critical public health concern. Predicting one's risk of heart disease is a difficult task owing to the intricate interaction of genetic, behavioural, and environmental variables. As new technology, particularly in the field of machine learning, advances, an increasing number of individuals are becoming interested in employing computational approaches to increase the accuracy and efficacy of cardiac disease prediction. Machine learning is a kind of artificial intelligence that allows computers to detect patterns and predict outcomes from data without requiring explicit programming. Machine learning algorithms can analyze big datasets including a range of patient variables, such as medical history, lifestyle choices, and genetic predispositions, in the context of predicting cardiovascular disease. These algorithms may assist medical workers better predict and prevent heart illness by identifying hidden patterns and relationships in this data. Predictive modelling has the ability to alter traditional risk assessment methodologies by providing personalized and data-driven insights. Unlike conventional risk calculators, which rely on a limited number of factors, machine learning models may consider a huge number of variables and respond to new data, enhancing prediction accuracy. Predictive modelling is an effective methodological tool that assists users in navigating the complicated data analytics environment by generating predictions about the future based on previous trends and patterns. Simply described, it is the process of developing and implementing algorithms to predict or identify potential patterns in data sets. This strategy is particularly useful in sectors where understanding and forecasting future occurrences are critical. Predictive modelling is a cutting-edge technique that may be used in marketing to target consumer behaviour, healthcare to anticipate the progression of a disease, and finance to forecast market trends. It uses statistical and mathematical approaches to transform raw data into actionable insights, allowing decision-makers to take proactive steps to handle challenges and capitalize on opportunities. Predictive modelling integration becomes more than just a tool; it becomes a revolutionary force as industries increasingly rely on data to impact strategies, allowing businesses to navigate complexity and make well-informed choices in an ever-changing environment. The value of reviewing a patient's entire medical history is becoming more obvious as healthcare advances, emphasizing its role as a critical tool in the pursuit of the greatest possible health outcomes.

2.LITERATURE REVIEW:

2.1 MACHINE LEARNING-BASED PREDICTION OF HEART DISEASE

Sean C. [1] The use of machine learning in the healthcare business has enormous promise, particularly for early detection and prognosis of various medical conditions. The application of machine learning algorithms is very significant in terms of heart health. Predicting potential cardiac problems ahead of time provides considerable advantages in terms of prompt treatment and personalized treatment approaches. The purpose of this research project is to examine and evaluate how well different machine learning classifiers perform in terms of heart condition prediction. Decision trees, Naive Bayes, Logistic Regression, Support Vector Machines (SVM), and Random Forest are among the classifiers under consideration. To establish which of these classifiers works best in the specific context of heart health prediction, a comparative study is required. Each of these classifiers brings unique strengths and characteristics to the table. In addition, the study proposes an ensemble classifier, a unique approach. This classifier incorporates the best aspects of both strong and weak classifiers, going beyond the conventional single-model approach. The idea for this hybrid classification technique is based on its ability to efficiently employ a large number of training and validation data. The ensemble classifier's goal is to increase the model's overall resilience and prediction accuracy, making it a more reliable tool for early identification of potential cardiac issues.

2.2 APPLICATION OF HYBRID MACHINE LEARNING TECHNIQUES FOR EFFECTIVE PREDICTION OF HEART DISEASE

SENTHILKUMAR MOHAN [2] Heart disease is a major cause of mortality and a worldwide public health problem. Predicting cardiovascular diseases is an important aspect of clinical data analysis. The large quantity of data generated by the healthcare industry necessitates new technology, and machine learning has shown to be very effective in this regard. Combining machine learning approaches with clinical data offers a promising approach to improving prediction accuracy and decision-making in cardiovascular health. The combination of machine learning and the Internet of Things (IoT) has transformed healthcare analytics. Recent breakthroughs have shown that integrating machine learning algorithms with IoT devices may provide real-time data that can help anticipate and prevent cardiac disease. This confluence of technology has increased the possibility for proactive and tailored healthcare treatments. Despite the amount of earlier research in the domain, the purpose of this study is to enhance the field of machine learning-based cardiac illness prediction by presenting a novel technique. The primary objective is to apply powerful machine learning algorithms to identify and harness significant features, thereby increasing the accuracy of cardiovascular disease predictions.

2.3 USING MACHINE LEARNING ALGORITHMS FOR PREDICTING HEART DISEASE

Shu Jiang [3]. Examining the consequences of cardiovascular diseases (CVDs) on a global scale demonstrates that this health condition affects a large number of people and is the leading cause of death worldwide, exceeding all other causes. According to the World Health Organization, CVDs caused 17.9 million deaths globally in 2016, accounting for 31% of total fatalities. Heart attacks and strokes accounted for 85% of the deaths. Given the high fatality rates and high cost of cardiovascular surgery, this bleak reality not only imposes a huge emotional load on the affected families, but it also causes enormous financial difficulties. Heart disease is a major and even uncontrollable danger in economically disadvantaged communities, where the situation is particularly dire. As a result, it becomes vital to investigate the intricate relationships between many human qualities and the risk of developing heart disease. Developing a solid predictive model is not simply a statistical challenge, but also an important tool for predicting and preventing cardiac diseases. Within this paradigm, machine learning technologies pitch themselves as an effective weapon against heart disease. Machine learning, which is closely connected to computational statistics, use mathematical optimization to create theories and techniques for addressing real problems in commerce, industry, social sciences, and medicine. The two major fields of machine learning, supervised learning and unsupervised learning, illustrate the field's variety. Supervised learning appears as the apparent answer to the specific goal of predicting heart disease based on physiological features.

2.4 USING RELIEF AND LASSO FEATURE SELECTION TECHNIQUES IN MACHINE LEARNING ALGORITHMS, PRANAB GHOSH [4] SUCCESSFULLY PREDICTS CARDIOVASCULAR DISEASE.

Cardiovascular diseases (CVDs) are a significant worldwide health concern owing to their widespread and negative impact on human health. Early detection has the potential to prevent or mitigate the impact of CVDs, thus identifying risk factors should be a major focus. Using machine learning models to predict heart illness seems to be a viable method in this case. The model proposed in this study aims to increase the accuracy of such predictions by integrating multiple methodologies. The effectiveness of the proposed model requires a rigorous approach to data management, which includes good data collecting, pre-processing, and transformation procedures. These activities are required to ensure the generation of exact and reliable data for the model's training. The incorporation of several datasets from Stat log, Long Beach VA, Cleveland, Switzerland, and Hungary adds to the model's comprehensiveness. This permits a wide range of data to be collected for analysis. Feature selection is an important step in increasing the model's predictive ability. To locate and choose the most relevant characteristics for this research, the Relief method and the Least Absolute Shrinkage and Selection Operator (LASSO) are utilized. The use of a strategic selection approach enhances the model's ability to detect and rank risk factors for heart disease. The integration of new hybrid classifiers, such as the Gradient Boosting Method (GBBM), AdaBoost Boosting Method (ABBM), K-Nearest Neighbours Bagging Method (KNNBM), Decision Tree Bagging Method (DTBM), and Random Forest Bagging Method (RFBM), is the study's innovative component. During training, these hybrid classifiers integrate bagging and boosting approaches with traditional classifiers.

2.5 PROJECTED HEART DISEASE HARSHIT JINDAL

states that with the frequency of cardiac problems on the increase, it is more crucial than ever to predict and recognize these illnesses early on. Because this diagnostic activity is so complex, efficiency and precision are required, necessitating the development of innovative techniques. The research paper being reviewed focuses on identifying people who are more likely to suffer heart disease based on a range of medical criteria. To tackle this daunting issue, the researchers developed a heart disease prediction algorithm that draws on individuals detailed medical histories. By forecasting a patient's propensity to heart disease, this method hopes to provide a pre-emptive approach to intervention. The use of a range of machine learning techniques, such as K-nearest neighbours (KNN) and logistic regression, demonstrates the versatility of modern computational approaches in medical diagnosis. Improving the accuracy of heart disease projections is an important part of the study. The authors carefully tweaked the model to optimize reliability and performance.

3.EXISTING SYSTEM

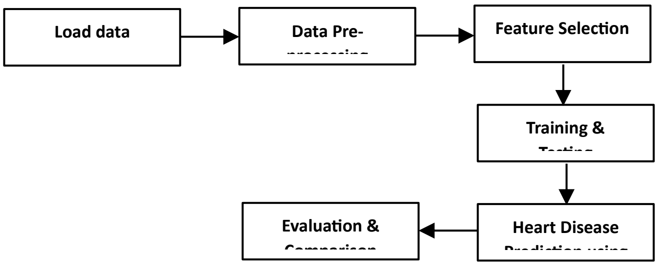
One of the most challenging jobs in medicine is to forecast cardiac disease. Doctors and other medical experts, in particular, spend a significant amount of time and effort identifying the source of this. This study predicts cardiac sickness using GridSearchCV and a variety of machine learning methods such as LR, KNN, SVM, and GBC. The system uses a 5-fold cross-validation approach to verify its results. A comparison of these four methodologies is shown. The models' performance is evaluated using datasets from Cleveland, Hungary, Switzerland, Long Beach V, and UCI Kaggle. The analysis demonstrates that, for both datasets (Hungary, Switzerland & Long Beach V and UCI Kaggle), the Extreme Gradient Boosting Classifier with GridSearchCV achieves the greatest and almost identical testing and training accuracies of 100% and 99.03%. Furthermore, the study demonstrates that, for both datasets (Hungary, Switzerland, Long Beach V, and UCI Kaggle), the XG Boost Classifier without GridSearchCV achieves the greatest and almost similar testing and training accuracies.

3.1 DRAWBACKS OF EXISTING METHOD:

Heart disease datasets often exhibit imbalanced classes (e.g., more healthy individuals than those with heart disease), leading to models that favor the majority class and struggle to accurately predict the minority class (those with heart disease).

3.2 THE SYSTEM PROPOSED

The proposed approach, Warm and Naive Bayes (NB), combines machine learning methods to develop a robust framework for forecasting cardiac disease. Using a huge dataset of essential health indicators, the system carefully preprocesses data, corrects outliers and missing values, and normalizes numerical properties. Predetermined metrics are used to train the models and evaluate their prediction ability. With an emphasis on deployment readiness, the proposed system aims to guarantee the proper use of health data throughout the process while also providing an accurate heart disease prognostic model and taking ethical factors into account. This comprehensive technique, which has the potential for broader applications in real-world settings, intends to enable the development of ethical and effective predictive healthcare solutions for heart disease



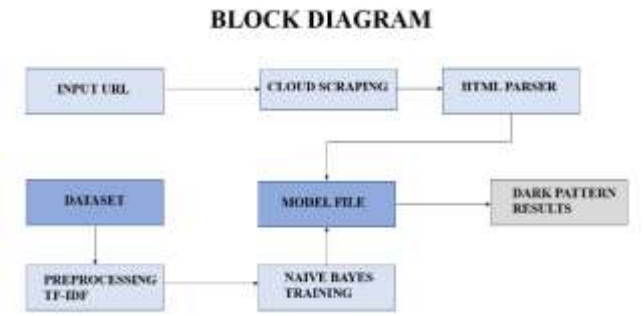


Fig 1. Block diagram for Heart Disease Prediction

BLOCK DIAGRAM DESCRIPTION: The heart disease prediction process begins with **Data Collection & Loading**, where key health factors like age, gender, blood pressure, and cholesterol are gathered to form the foundation for model development. Next, **Data Preprocessing** ensures data integrity by handling missing values and outliers through techniques like imputation and normalization. **Feature Selection** then identifies the most relevant factors to enhance model accuracy and reduce complexity. The dataset is then **split into Training & Testing** sets, where models learn patterns and are evaluated on unseen data. Finally, **Prediction Using Warm & Naïve Bayes** applies these algorithms to assess heart disease risk, allowing comparison to determine the most effective model.

ADVANTAGES OF PROPOSED SYSTEM:

The proposed approach integrates machine learning methods, particularly Warm and Naïve Bayes (NB), to create a robust heart disease prediction framework. It meticulously preprocesses a vast array of critical health indicators, addressing missing values, outliers, and standardizing numerical features to ensure data consistency and reliability.

4.SYSTEM IMPLEMENTATION

Module 1: Load Data:

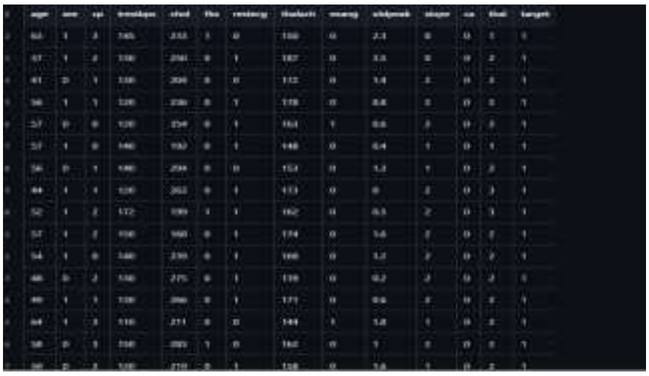
1 Acquire and load the heart disease dataset containing critical health parameters, including age, gender, blood pressure, cholesterol levels, and lifestyle factors. Ensure data quality by verifying completeness, consistency, and accuracy. Store the dataset in a structured format suitable for preprocessing and analysis

PROGRAM: DATA COLLECTION

Fig 2: Program of Data Collection for Heart Disease Prediction Using Machine Learning.

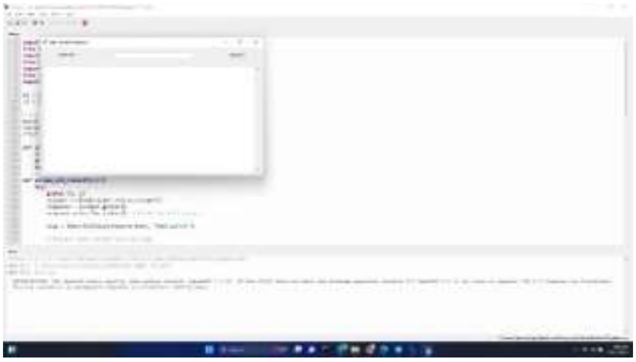
Module 2: Data Preprocessing

This module ensures the trustworthiness of the dataset by addressing issues like outliers and missing values.



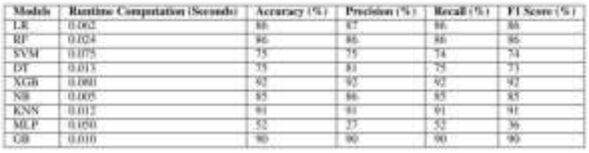
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
10	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
11	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
12	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
13	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
14	58	0	3	150	283	1	0	162	0	1	2	0	2	1
15	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1

While scaling techniques normalize numerical features for consistency, imputation and normalization techniques are used to preserve data integrity. In order to provide a clear and organized basis for further analysis, duplicate records are also located and eliminated, and categorical variables are encoded for machine learning algorithm.

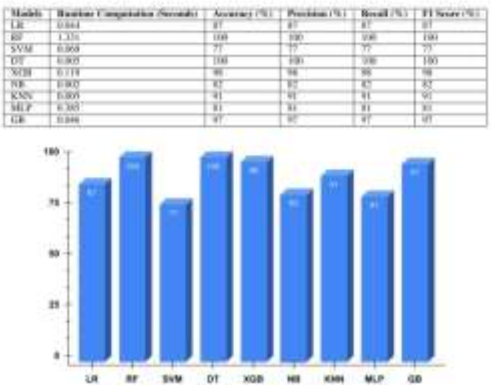


Module 3: Feature Selection

Feature selection is a key technique for identifying the most relevant attributes in a dataset that significantly influence disease outcomes. In high-dimensional datasets, selecting the most crucial features enhances the accuracy of prediction models and helps reduce the likelihood of medical misdiagnoses. The selection process is based on ranking features according to their importance. To achieve this, the authors utilized the Latent Feature Selection (ILFS) method, a probabilistic latent graph-based approach that efficiently ranks features based on their predictive impact.



Models	Runtime Computation (Seconds)	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
LR	0.062	86	87	86	86
RF	0.024	86	86	86	86
SVM	0.075	75	75	74	74
DT	0.013	75	81	75	73
XGB	0.080	92	92	92	92
NB	0.005	85	86	85	85
KNN	0.012	91	91	91	91
MLP	0.050	52	27	52	36
GB	0.010	90	90	90	90



Models	Runtime Computation (Seconds)	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
LR	0.044	87	87	87	87
RF	1.331	100	100	100	100
SVM	0.068	77	77	77	77
DT	0.005	100	100	100	100
XGB	0.119	98	98	98	98
NB	0.002	82	82	82	82
KNN	0.005	91	91	91	91
MLP	0.305	81	81	81	81
GB	0.046	97	97	97	97

Fig 2: Feature Selection for Heart Disease Prediction Using Machine Learning

Module 4: Training and Testing Data

The **training set** is used to train the machine learning model by identifying patterns and understanding relationships between features and outcomes. It typically comprises a larger portion of the dataset (around 70–80%) to ensure comprehensive learning. To enhance performance, the training set can be further split into validation data for fine-tuning hyperparameters, and data augmentation techniques can be applied in cases like image or text processing. A well-structured training set helps prevent underfitting by providing sufficient data for the model to generalize effectively.

The **testing set**, which usually makes up 20–30% of the dataset, is used to evaluate the model's ability to handle new, unseen data. This step ensures generalization and helps detect overfitting if the model performs well on training data but poorly on testing data. The testing set also aids in comparing different models, selecting the best-performing one, and validating the effectiveness of the learning process. Additionally, key performance metrics such as accuracy, precision, recall, and F1-score are calculated based on the testing set to measure the model's predictive capability in real-world applications.

Module 5: Heart Disease Prediction using Naive Bayes and Warmth Algorithm

1.Warmth Algorithm- WARMTH analyses relationships within the data, assigning weights to different features based on their relevance to the target outcome (heart disease), effectively highlighting the most impactful factors in the prediction.

2. Naive Bayes: Once the relevant features are identified by WARM, the Naive Bayes classifier calculates the probability of a patient having heart disease based on their values for those features, making a prediction based on the most likely outcome.

RESULT ANALYSIS

Three models were examined in terms of accuracy in predicting algorithms for the detection of heart disease. With an impressive 75% accuracy score, the XGBoost Classifier exhibited its ability to grasp the dataset's deep linkages. Warm and Naive Bayes algorithms outperformed XGBoost, with an incredible 88% accuracy. This considerable accuracy demonstrates that the Warm and Naive Bayes models are effective predictors when applied with a diverse dataset that contains crucial health parameters such as age, gender, blood pressure, and cholesterol levels. The huge 13% improvement in accuracy over XGBoost demonstrates how effective Bayesian approaches are in predicting cardiac disease. These findings demonstrate how Warm and Naive Bayes algorithms may increase prediction model accuracy in healthcare applications, particularly in heart disease prognosis.

Train and Compare Classifiers

Decision Tree Metrics

```
{
  "Accuracy" : [illegible]
  "Precision" : [illegible]
  "Recall" : [illegible]
  "F1 Score" : [illegible]
}
```

Naive Bayes Metrics

```
{
  "Accuracy" : [illegible]
  "Precision" : [illegible]
  "Recall" : [illegible]
  "F1 Score" : [illegible]
}
```

Model Comparison

Fig 3:Accuraccy Prediction

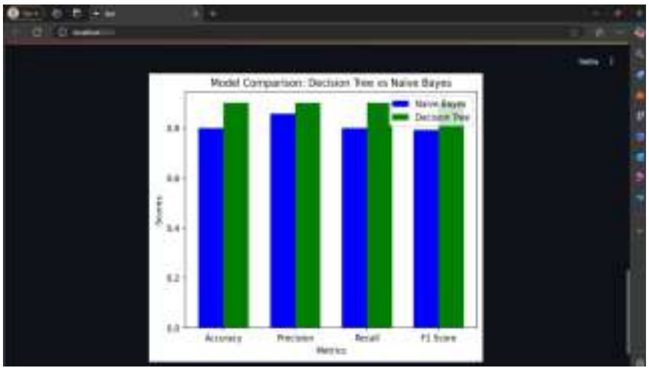


Fig 4: Decision Tree vs Naive Bayes

5.FUTURE WORK

Subsequent research in this area may look at the use of advanced feature engineering approaches and deep learning structures to increase the predictive potential of heart disease models. By examining the impacts of numerous health markers and incorporating real-time monitoring data, we may get a better knowledge of the dynamic nature of cardiovascular health. Furthermore, efforts must be directed toward developing interpretable models in order to increase predictability and transparency, especially in instances when critical healthcare choices must be made.

6.CONCLUSION

To summarize, this study established a comprehensive framework for predicting cardiac illness using machine learning approaches like as Warm and Naive Bayes (NB). The system's purpose is to create accurate and trustworthy predictions based on significant health parameters via thorough data preprocessing, feature selection, and model training. Algorithm comparisons provide light on each's merits and weaknesses, providing critical information for making an educated model decision. Throughout the process, ethical considerations have been critical to ensuring the proper use of health data**.**

REFERENCE

[1] K. T. and Agarwal. Kumar, at the 2nd International Conference on Intelligent Computing and Control Systems (ICICCS), "heart disease prediction using machine learning." Ieee, 2018

[2] S. Rajput and A. Arora, "Hybrid machine learning techniques for effective prediction of heart disease," International Journal of Computer Applications, vol. 2013, 75, no. 10, pp. 6–12.

[3] M. A. and Mohammed Selamat, "Machine learning algorithms for heart disease prediction," International Conference on Computer, Communications, and Control Technology (i4ct), 2015, IEEEE, pp. 227-231

[4] In the proceedings of the first instructional conference on machine learning, vol., Ramos et al. describe "efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques." 242. 133–142 in Piscataway, New Jersey, 2003.

[5] T. Kumaresan and C. Palanisamy, "Prognostication of Heart Disease," International Journal of Bio-inspired Computing, vol. 9, no. 3, 2017, pp. 142-156.

[6] H.S. and Kaur. "Improving the accuracy of heart disease prediction using machine learning methods and optimization," Ajay, Next Generation Computing Technologies (ngct), 2016, pp. 516–521.

[7] K. Toutanova and C. Cherry, "Heart disease prediction using machine learning and svm techniques," in Proceedings of the 4th International Joint Conference on Natural Language Processing of the American Foundation for Nursing and Palliative Care, Volume 1-Vol. 1, 2009, pp. 486-494, Association for Computational Linguistics.

[8]]. T. Sainath, N. O. Vinyals, and a senior. Sak, "Optimization of energy consumption in container-oriented cloud computing centers," Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (icassp). IEEEE, 2015, 4580–4584 pages

[9] T. Mikolov and G. Zweig, "Machine learning for cardiovascular diseases (CVDs)," at the 2012 IEEE Spoken Language Technology Workshop. Pages 234-239 of IEEEE, 2012.

[10] Rizky, W. M., Afrizal, D., and Ristu, S. "Heart disease prediction system using sequential backward selection algorithm for features selection and machine learning model." Journal of Scientific Informatics, Volume 3(2), November 2020, Pages 41-50.