MACHINE LEARNING BASED PATIENT CLASSIFICATION IN EMERGENCY DEPARTMENT

E.Dileep[1],G.Sreelekha[2], K.Nandhini[3],P.Sreekanth[4],T,Karthik[5], T.Prameela[6]

[1,2,3,4,5] UG Student, Dept. of CSE, Siddartha Educational Academy Group of Institutions, Tirupati AP, India
[6] Professor, Dept. of CSE, Siddartha Educational Academy Group of Institutions, Tirupati, AP, India

Abstract- This work contains the classification of patients in an Emergency Department in a hospital according to their critical conditions. Machine learning can be applied based on the patient's condition to quickly determine if the patient requires urgent medical intervention from the clinicians or not. Basic vital signs like Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Respiratory Rate (RR), Oxygen saturation (SPO2), Random Blood Sugar (RBS), Temperature, Pulse Rate (PR) are used as the input for the patients' risk level identification. High-risk or non-risk categories are considered as the output for patient classification. Basic machine learning techniques such as LR, Gaussian NB, SVM, KNN and DT are used for the classification. Precision, recall, and F1-score are considered for the evaluation. The decision tree gives best F1-score of 77.67 for the risk level classification of the imbalanced dataset.

Keywords: Machine Learning, Triage, Numpy, Pandas, Logistic regression, Decision tree, SVM.

1.Introduction

In the Emergency Departments of hospitals, patients are sorted based on their need for immediate medical treatment. This sorting is done according to the urgency or severity of the health conditions of patients. When a patient arrives, an ER (emergency room) nurse performs a brief, focused assessment and assigns the patient a triage acuity level, also known as a triage score. Triage [1] establishes priorities for care and determines the clinical area of treatment. The acuity level is a proxy measure of how long the patient can safely wait for medical evaluation and treatment. For this purpose, healthcare workers categorize them as per their risk level. Priority level 1 patients are critically ill or high-risk category patients and need immediate medical attention to save their life. This is done by nurses or the assigned staff at hospital triage considering their vital signs and clinical observations. Priority level 2 patients are those who need medical attention but can wait as long as 30 minutes for assessment and treatment. These patients are considered medium-risk level patients. Other cases are considered low-risk patients. They can wait for medical help. This type of patient classification is done by considering their basic vital signs and clinical conditions.

This work focuses on machine learning algorithms to automatically classify critical and non-critical patients based on measured signs. Machine Learning algorithms

Triage is the prioritization of injured or sick individuals based on their need for emergency treatment. Each organization will have its own triage system, which often includes color coded categories. Triage may be used to meet an organization's short or long-term needs to help determine who gets care first. Based on these results, Machine Learning can determine the patient's criticality. In this study, basic vital parameters are used for patient classification as input. Medium-risk patients and low-risk patients are considered non-critical patients while high-risk cases are considered critical patients. The vital parameters used are Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Respiratory Rate (RR), Oxygen saturation (SPO2), Random Blood Sugar (RBS), Temperature, and Pulse Rate (PR). The output is taken as the patient whose condition is critical falls under class 1 and non-critical patients are classified under class 0.

This work focuses on machine learning algorithms to automatically classify critical and non-critical patients based on measured signs. Machine Learning algorithms executed in this work are Gaussian Na¨ıve Bayes Classifier (NBC), Logistic Regression (LR), Support Vector Machine (SVM), K Nearest Neighbours (KNN), and Decision Trees (DTs). All obtained results are

compared to evaluate the effectiveness of each method.

The remaining part of the paper is organized as follows: Section II is Literature Survey describing already existing works. Section III is Methodology which highlights the dataset being worked on and the proposed algorithms. Section IV describes the Experimental results that are obtained after building classifiers and discusses the performance evaluation of all classifiers. Section V is the Conclusion that concludes the overall work

II.Literature Review

Sl.No.	Name of the Aurthor(s)	Title	Description	Year
1	Tintinalli, Judith E	Disaster Preparedness", Tintinalli's Emergency Medicine: A Comprehensive Study Guide, 9th Edition, McGraw-Hill Education	Disasters have claimed millions of lives and cost billions of dollars worldwide in the past few decades. Emergency physicians frequently have extensive responsibilities for community and hospital-level disaster preparedness and response	2019
2	D.A Debal, T.M. Sitote	Chronic kidney disease prediction using machine learning techniques", Journal of Big Data	Goal three of the UN's Sustainable Development Goal is good health and well-being where it clearly emphasized that non-communicable diseases is emerging challenge. One of the objectives is to reduce premature mortality from non-communicable disease	2022

3	K.M. Almustafa	Prediction of chronic kidney disease using different classification algorithms	Diabetes mellitus is a serious health issue in healthcare industry, which is a type of uncontrolled level of sugar. In the predictive system, feature selection plays on vital role to select the relevant feature for classification. There are several algorithms were applied on classification of diabetes data. In this proposed work, the features are transformed into high dimensional space before selection.	2021
4	T. Nibareke, J. Laassiri	Using Big Data-machine learning models for diabetes prediction and flight delays analytics Journal of Big Data	Large data volumes are daily generated at a high rate. Data from health system, social network, financial, government, marketing, bank transactions as well as the censors and smart devices are increasing. We highlight some metrics that allow us to choose a more accurate model. We predict diabetes disease using three machine learning models and then compared their performance. Further more we analyzed flight delay and produced a dashboard which can help managers of flight companies to have a 360° view of their flights and take strategic decisions.	2020

III.Methodology

Required Algorithms

The algorithms used in the provided code are:

1. Random Forest Classifier
2. Decision Tree Classifier
3. Support Vector Machine (SVM)
4. K-Nearest Neighbours (KNN) Classifier
5. Logistic Regression Classifier

Explanation of Algorithms

1. Random Forest Classifier

Random Forest is a powerful and versatile machine learning algorithm that has gained widespread popularity for its effectiveness in both classification and regression tasks. It belongs to the ensemble learning category, which involves combining the predictions of multiple models to enhance overall performance. Introduced by Leo Bierman in 2001, the Random Forest algorithm is known for its ability to handle complex datasets and deliver robust results. Classification proposed by Eugene Kleinberg.

Firstly, the algorithm starts with the process of bootstrapped sampling. Multiple random subsets, called bootstrap samples, are generated by randomly selecting instances with replacement from the original training dataset. This step introduces diversity among the subsets, ensuring that each decision tree in the ensemble is trained on a slightly different variation of the data.

Next, the construction of decision trees takes place. For each bootstrap sample, a decision tree is grown by recursively partitioning the data based on randomly selected features at each node. The randomness in feature selection prevents individual trees from becoming overly specialized to certain features, reducing the risk of overfitting. The trees are grown until a specified criterion is met, such as a predefined depth or the inability to further split the data.

INPUT AND OUTPUT SPECIFICATIONS

Input Requirements

The Random Forest algorithm, like many machine learning algorithms, requires specific input data to train and make predictions. The key input requirements for the Random Forest algorithm include:

Feature Matrix:

The primary input is a feature matrix that represents the training data. Each row of the matrix corresponds to an instance or observation, and each column corresponds to a feature or attribute. The matrix should be numeric and can include both continuous and categorical features.

Target Variable (Response Variable):

For supervised learning tasks (such as classification or regression), there must be a target variable or response variable. This variable represents the outcome or label that the algorithm aims to predict. The target variable should be associated with each instance in the training data.

Training Data:

Random Forest requires a sufficiently large and representative training dataset. The dataset should cover a diverse range of scenarios and patterns that the algorithm can learn from.

Categorical Encoding (if applicable):

If the dataset includes categorical features, they need to be appropriately encoded. Random Forest can handle categorical variables, but they often need to be converted into numerical representations. Common methods include one-hot encoding or label encoding.

Imputation of Missing Values (if applicable):

If the dataset contains missing values, it's essential to handle them appropriately. Random Forest can tolerate missing values, but imputation methods such as mean imputation or more sophisticated techniques may be applied to address this issue.

Balanced Classes (for Classification):

In classification tasks, it is desirable to have a balanced distribution of classes in the training data. Extreme class imbalance might require additional techniques, such as class weighting or resampling, to ensure that the model is not biased toward the majority class.

Ensuring that the input data meets these requirements is crucial for the successful training and deployment of a Random Forest model. Proper preprocessing, handling of missing values, and careful consideration of the characteristics of the dataset contribute to the algorithm's effectiveness and generalization to

Output Format

The output of the Random Forest algorithm is multifaceted and includes various components that provide insights into the model's performance and predictions. The primary outputs are associated with both the training phase, where the model learns from the data, and the prediction phase, where it makes predictions on new or unseen instances.

Trained Random Forest Model:

The most fundamental output is the trained Random Forest model itself. This consists of an ensemble of decision trees, each capturing different aspects of the relationships within the training data. The model retains the knowledge gained during the training phase, encapsulating the patterns and decision rules learned from the input features and target variable.

Prediction Results:

When the trained Random Forest model is applied to new data, it produces predictions. For classification tasks, the output typically includes predicted class labels for each instance. For regression tasks, the output consists of predicted numerical values. These predictions are generated by aggregating the individual predictions from each tree in the ensemble.

Applications:

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression. Many other medical scales used to assess severity of a patient have been developed using logistic regression Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.) Another example might be to predict whether a Nepalese voter will vote Nepali Congress or Communist Party of Nepal or Any Other Party, based on age, income, sex, race, state of residence, votes in previous elections, etc. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription, etc in economics, it can be used to predict the likelihood of a person ending up in the labour force, and a business application would be to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing.

Logistic regression finds extensive applications across various domains due to its versatility, simplicity, and interpretability. One prominent area of application is in **Healthcare and Medicine**.

Logistic regression is frequently employed to predict the likelihood of a patient having a particular medical condition based on various risk factors. For instance, it can be used to predict the probability of a patient developing a specific disease, such as diabetes or heart disease, considering factors like age, BMI, and family medical history. In epidemiology, logistic regression is utilized to analyse and model the risk factors associated with the occurrence of diseases within populations.

In **Marketing and Business**, logistic regression plays a crucial role in customer churn prediction. Companies leverage this algorithm to assess the probability of a customer discontinuing their services based on factors like usage patterns, customer feedback, and billing history. Additionally, logistic regression is employed in credit scoring models to evaluate the risk of default for loan applicants. By considering variables such as credit score, income, and debt-to-income ratio, financial institutions can make informed decisions about whether to approve or deny a loan application.

Within the **Social Sciences**, logistic regression is widely used for behavioural analysis and predicting outcomes. Sociologists may use logistic regression to understand factors influencing voting behaviour, predicting the likelihood of an individual participating in a particular social activity, or even exploring the probability of someone adopting a certain behaviour or attitude. In psychology, logistic regression can be applied to model the probability of an individual belonging to a specific psychological profile based on various personality traits or environmental factors.

In the realm of **Information Technology and Cybersecurity**, logistic regression is employed for intrusion detection systems. By analysing network traffic patterns, user behaviours, and system logs, logistic regression models can identify abnormal activities and predict the likelihood of a security breach. This aids inearly detection and prevention of cyber threats. Moreover, logistic regression is utilized in sentiment analysis of user reviews, helping businesses understand customer opinions about products or services by predicting whether a review is positive or negative.

Logistic regression is also extensively used in **Environmental Science and Ecology**. For example, it can be applied to model the probability of species occurrence or habitat suitability based on environmental variables like temperature, precipitation, and land cover. Conservationists use logistic regression to assess the impact of various factors on endangered species' survival probabilities, aiding in the development of effective conservation strategies.

In **Finance and Economics**, logistic regression is applied to predict events such as bankruptcy, stock price movements, or the success of a financial product. For credit risk assessment, logistic regression models are used to evaluate the probability of a borrower defaulting on a loan. In econometrics, logistic regression helps analyse the impact of independent variables on the probability of an economic event, such as the likelihood of a company going public.

In **Education and Educational Research**, logistic regression is employed to study factors affecting educational outcomes. Researchers may use logistic regression to predict the probability of student success or dropout based on variables like socioeconomic status, previous academic performance, and attendance. This aids in identifying interventions to support students at risk.

In conclusion, logistic regression's broad applicability makes it a valuable tool across diverse fields, contributing to informed decision-making and predictive modelling in areas ranging from healthcare and finance to social sciences and environmental studies.

Working

Emergency Department using machine learning typically involves the following steps:

Data Collection: The data collection used to classify results for the better classification of patients' risk conditions. In this work, primary data are collected from the Emergency
Medicine Department of a leading multi specialitsy hospital, Kerala, India. Basic emergency vitals are collected with a criticality index, which is the target value. 2578 cases are noted from the hospital, leaving patients' personal information
behind. 519 cases are in critical condition out of 2578 cases

1. **Data Preprocessing**: The collected data undergoes preprocessing to handle missing values, outliers, and inconsistencies. This step may involve techniques such as imputation,outlier detection, normalization, and feature scaling to ensure the quality and consistency of the data.

2. **Feature Selection and Engineering**: Features that have a significant impact on groundwater levels are identified and selected for model training. Additionally, new features may be created through feature engineering to capture complex relationships or temporal dependencies in the data.

3. **Model Training**: Machine learning algorithms such as regression models, decision trees, support vector machines, or ensemble methods are trained on the pre-processed data to learn the underlying patterns and relationships between the input features and patient dataset. The dataset is typically split into training and testing sets to evaluate the performance of the trained models.

4. **Model Evaluation**: The performance of the trained models is evaluated using appropriate evaluation metrics such as root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R-squared), or others, depending on the specific requirements of the application. Cross-validation techniques may also be employed to ensure the robustness and generalization capability of the models.

5. **Model Selection and Optimization**: Different machine learning algorithms and model configurations are compared based on their performance metrics, and the best-performing model is selected for further optimization if necessary. Hyperparameter tuning techniques such as grid search or random search may be used to fine-tune the model parameters for improved performance.

   Trained models are rigorously evaluated using appropriate performance metrics, such as root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R-squared), and Nash-Sutcliffe efficiency (NSE). Cross-validation techniques, such as k-fold cross-validation or time-series splitting, ensure the generalizability of the models across different temporal and spatial contexts. Model validation involves comparing predicted groundwater levels against observed measurements to assess the model's accuracy and reliability.

6. **Prediction and Deployment**: Once the model is trained and evaluated satisfactorily, it can be used to make predictions on new or unseen data. The trained model can be deployed as part of an operational system for real-time monitoring and prediction of risk levels, providing valuable insights for decision-making in emergency department.

Overall results for the better classification of patients' risk conditions. In this work, primary data are collected from the Emergency Medicine Department of a leading multi speciality hospital, Kerala, India. Basic emergency vitals are collected with a criticality index, which is the target value. 2578 cases are noted from the hospital, leaving patients' personal information behind. 519 cases are in critical condition out of 2578 cases. Figure 1 indicates the feature distribution of Triage Vital Dataset.

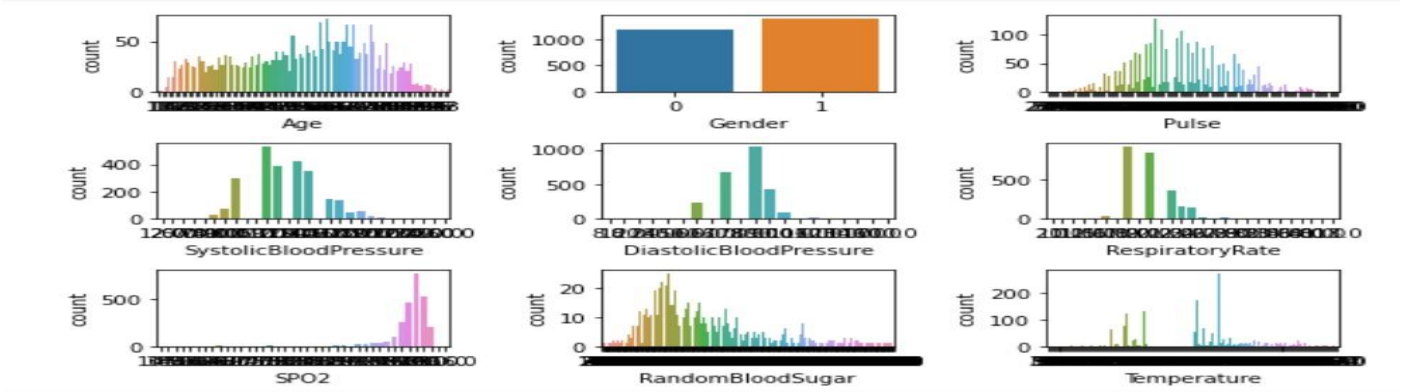

Fig 1

The proposed system employs various machine learning algorithms to build a system that classifies patients into different risk categories based on their basic vital statistics. We have considered 2 classes according to the patients risk level as Class 1: high risk level and class 0: low risk level. Out of 11 features in this Triage Vital Dataset, 7 features have missing values. Features with missing values and their corresponding numbers are shown in table I. For pre-processing, we used a statistical imputation method - mean value for filling missing values in Triage Vital Dataset.

TABLE I
MISSING VALES DESCRIPTION OF TRIAGE VITAL DATASET

Feature name	No.of missing values
Pulse	14
Systolic blood pressure	9
Diastolic blood pressure	11
Respiratory rate	22
SPO2	24
Random blood sugar	660
Temperature	40

Figure 2 shows the Pearson-ranking visualization of the Triage Vital Dataset after filling the missing values. A popular method Standard scalar is used to standardize the features, for data normalization in this work. Standard scalar removes means and scales the data into the unit variance.

Gini is used as a criterion for the decision tree, with a maximum depth of 5. The assumed random state is 33. In KNN, the number of neighbours is set to 5, each with uniform weight, and an auto algorithm is used. For the LR algorithm, the maximum iteration is taken as 1000 with C=1, and the l2 penalty is used along with the liblinear solver. Also in this case, the random state is taken as 33.

The machine learning approaches used in this work are Gaussian Na¨ıve Bayes Classifier (NBC), Logistic Regression (LR), Support Vector Machine (SVM), K Nearest Neighbours (KNN) and Decision Tree (DT) for the classification of patients. For better performance of these algorithms with the triage vital dataset, the hyperparameter settings are utilized in each model are shown in table II.

TABLE II
HYPER PARAMETER SETTINGS FOR CLASSIFICATION MODELS

Model	Hyper parameter settings
LR	penalty =’l2’, solver =’liblinear’, C=1, max iter=1000, random state=33
Gaussian NB	BernoulliNB
KNN	n neighbors=5, weights = ’uniform’, algorithm=’auto’
SVM	kernel= ’rbf’, max iter=4000, C=10, gamma=1
DT	criterion =’gini’, max depth=5, random state=33

monitoring and adaptive management strategies facilitated by predictive modeling enhance resilience to changing environmental conditions and anthropogenic influences.

However, it's essential to acknowledge the inherent uncertainties associated with groundwater level prediction, stemming from limitations in data

Logistic Regression, KNN and Gaussian Naive Bayes. These experimental findings used to evaluate the performance of all five classifiers and to suggest the algorithm that is most appropriate for classifying patients according to their risk level.

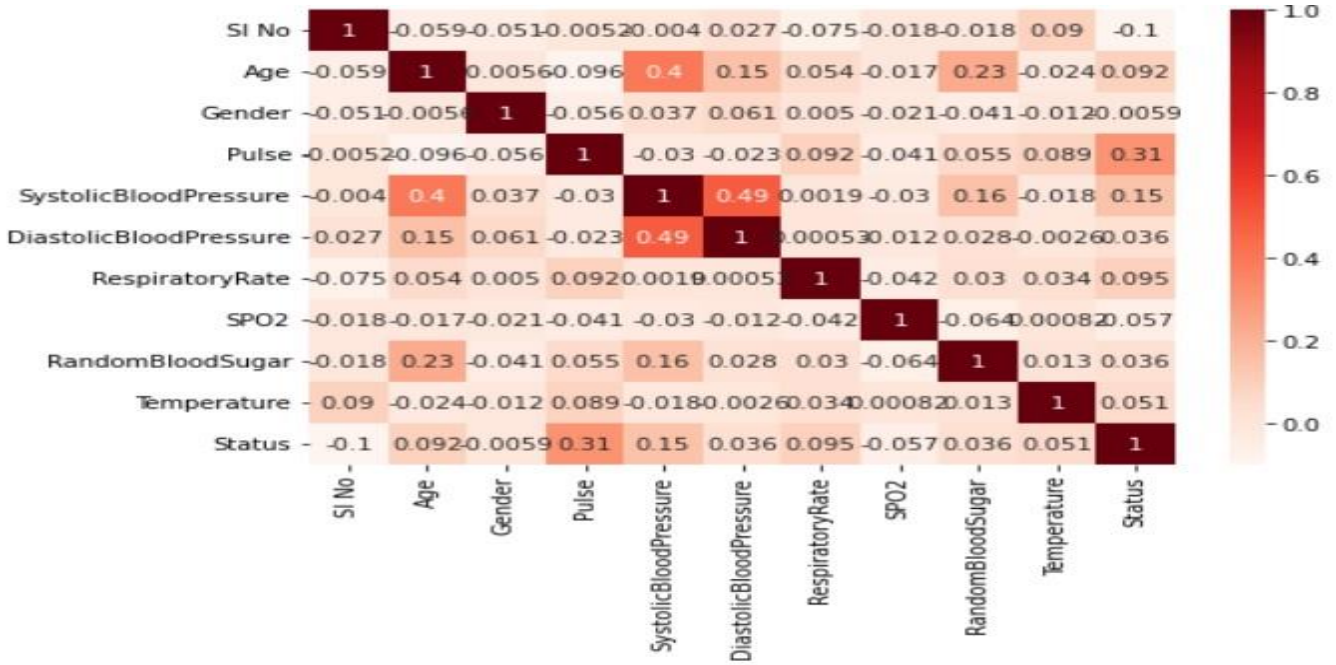

Fig. 2

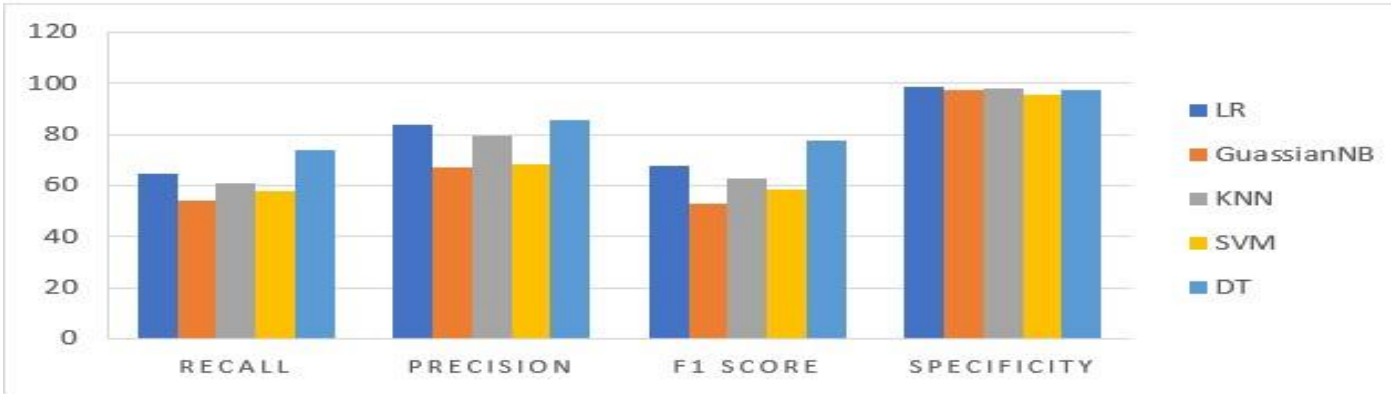

Fig.3

IV. RESULT AND DISCUSSIONS

This section outlines the experimental findings that were attained after the data set was trained and tested using classifiers such as the Support Vector Machine, Decision Tree, Logistic Regression, KNN and Gaussian Naive Bayes. These experimental findings used to evaluate the performance of all five classifiers and to suggest the algorithm that is most appropriate for classifying patients according to their risk level.

TABLE III

	LR	Gaussian NB	SVM	KNN	DT
Recall	64.38	54.02	60.70	57.49	74.03
Precision	83.58	66.95	79.39	68.32	85.56
F1 score	67.41	52.57	62.54	58.13	77.67
Specificity	98.34	97.51	98.01	95.69	97.18

In this work, we divided the Triage Vital Dataset into training and testing purposes with a ratio of 70 and 30 respectively. Recall, precision, and F1-score with specificity are the evaluation metrics for classification. Like any medical dataset, this Triage Vital Dataset is also an imbalanced dataset. Therefore, the F1-score is considered as the performance metric for the classification of patients' critical conditions. In this work, the classification is done with the above mentioned machine learning algorithms, and the hyperparameter settings are shown in Table II. The results are compared in table III
and figure 3 shows graphical representation of performance metrics. The decision tree classifier gave a maximum F1-score of 77.67 for this data, along with a better specificity of 97.18. LR gives the maximum specificity of 98.34 for the Triage Vital Dataset.

V. CONCLUSION

The healthcare industry can be benefited from machine learning especially in classification of patients condition. The algorithms proved that it is more useful in classifying the risk level of patients' conditions in triage as critical or noncritical. This system would help in reducing time delay to classify patients at triage in the Emergency Department of a hospital. This work proved that for the unbalanced Triage Vital Dataset, Decision Tree experimentally verified F1-score of 77.67 with a high specificity of 97.18. Moreover, this system can be useful in a pandemic situation in which all the resources are exhausted that happened during the Covid 19. This will maximize the efficiency of the health infrastructure and if the health industry takes this approach, it will reduce the workload of doctors and allow them to provide proper treatment to patients as soon as possible. Finally, it can be stated that the proposed system will benefit both doctors and patients.

VI.REFERENCES

[1] Tintinalli, Judith E, "Disaster Preparedness", Tintinalli's Emergency Medicine: A Comprehensive Study Guide, 9th Edition, McGraw-Hill Education, 2019, ISBN: 1260019934.

[2] D. Sharathchandra and M. R. Ram, "ML Based Interactive Disease Prediction Model," 2022 IEEE Delhi Section Conference (DELCON), 2022, pp. 1-5, doi: 10.1109/DELCON54057.2022.9752947.

[3] R. Krishnamoorthi, S. Joshi, H.Z. Almarzouki, P.K. Shukla, A. Rizwan, C. Kalpana, B. Tiwari, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", Journal of healthcare engineering, 1684017, Jan 2022 https://doi.org/10.1155/2022/1684017

[4] D.A Debal, T.M. Sitote, "Chronic kidney disease prediction using machine learning techniques", Journal of Big Data 9, Nov 2022, 10.1186/s40537-022-00657-5.

[5] K.M. Almustafa, "Prediction of chronic kidney disease using different classification algorithms", Informatics in Medicine Unlocked, 2021, Volume 24, 100631, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100631.

[6] A. R. Rao and B. S. Renuka, "A Machine Learning Approach to
Predict Thyroid Disease at Early Stages of Diagnosis," 2020 IEEE International Conference for Innovation in Technology

(INOCON), 2020, pp. 1-4, doi: 10.1109/INOCON50539.2020.9298252.

[7] T. Nibareke, J. Laassiri, “Using Big Data-machine learning models for diabetes prediction and flight delays analytics Journal of Big Data, 2020, 7, pp1-18 10.1186/s40537-020-00355-0

[8] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, “Comparing different supervised machine learning algorithms for disease prediction” BMC medical informatics and decision making, Dec 2019, 19(1), 281, doi:10.1186/s12911-019-1004-8

[9] T. T. Han, H. Y. Pham, D. S. L. Nguyen, Y. Iwata, T. T. Do, K. Ishibashi, G. Sun, “Machine learning based classification model for screening of infected patients using vital signs”, Informatics in Medicine Unlocked, Volume 24, 2021, 100592, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100592.

[10] M. Deepika and K. Kalaiselvi, “A Empirical study on Disease Diagnosis using Data Mining Techniques”, 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 615-620, doi: 10.1109/ICICCT.2018.8473185.

[11] A. Mir and S. N. Dhage, ”Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare,” 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697439.

[12] E. A. Choi et al., ”Prediction of COPD severity based on clinical data using Machine Learning,” 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 1646-1648, doi: 10.1109/BIBM52615.2021.9669887, ET-SIP-2254415.2022.9791739.