Patient Disease and Mortality prediction using EMRs

RUDRA KOSAMBI, KENIL GHETIA

Department of Computer Science, Pandit Deendayal Energy University

**ABSTRACT**

This research paper investigates the use of machine learning and artificial intelligence (AI) models for patient mortality and disease prediction from electronic medical records (EMRs) in healthcare. The paper aims to develop accurate predictive models for patient outcomes based on a comprehensive analysis of EMR data. The research methodology includes data preprocessing, feature selection, model selection, and evaluation. The study uses a large EMR dataset that includes information such as patient demographics, medical history, vital signs, lab results, diagnoses, and medications. Several machine learning and AI models are evaluated for mortality and disease prediction, including logistic regression, decision tree, random forest, support vector machine and k-nn classifier. The performance of each model is evaluated using metrics such as accuracy, sensitivity, specificity, and area under the curve. The results show that the artificial neural network model outperforms other models in terms of accuracy and sensitivity for both mortality and disease prediction. The study also discusses the ethical considerations and challenges associated with using machine learning and AI models in healthcare, such as data privacy and interpretability. The findings of this study can be used to develop predictive models that assist clinicians in making more informed decisions for patient care, ultimately leading to better patient outcomes.

1. **INTRODUCTION**

The increasing amount of electronic medical records (EMRs) available in healthcare has led to a growing interest in using machine learning and artificial intelligence (AI) models for patient mortality and disease prediction. EMRs contain a wealth of patient data that can be analyzed to identify risk factors and predictors for adverse outcomes, such as mortality and disease progression. The ability to accurately predict patient outcomes can assist clinicians in making more informed decisions for patient care, ultimately leading to better patient outcomes.

In this research paper, we investigate the use of machine learning and AI models for patient mortality and disease prediction from EMRs. We aim to develop accurate predictive models based on a comprehensive analysis of EMR data. The research methodology includes data preprocessing, feature selection, model selection, and evaluation.

We use a large EMR dataset from a hospital that includes information such as patient demographics, medical history, vital signs, lab results, diagnoses, and medications. We evaluate several machine learning and AI models for mortality and disease prediction, including logistic regression, decision tree, random forest, support vector machine, and artificial neural network. The performance of each model is evaluated using metrics such as accuracy, sensitivity, specificity, and area under the curve.

The findings of this research can be used to improve patient outcomes by assisting clinicians in making more informed decisions for patient care. However, the use of machine learning and AI models in healthcare also raises ethical considerations and challenges, such as data privacy and interpretability. Therefore, we also discuss the ethical considerations and challenges associated with using machine learning and AI models in healthcare.

1. **LITERATURE REVIEW**

Dhiraj Dahiwade et. al. [1] applied K-Nearest Neighbour (KNN) and Convolutional Neural Network (CNN) algorithms for disease prediction from symptoms and found that CNN is better algorithm in terms of accuracy and time with 84.5% accuracy.

S. Leoni Sharmila et. al. [2] performed comparative study of multiple machine learning techniques like Fuzzy logic, Fuzzy neural networks and decision tree classifier in classifying Liver disorders and found Fuzzy neural networks to be most effective model with accuracy of 91%.

Kohli P.S. et al. [3] applied multiple ML models on breast cancer, diabetes and heart disease datasets and found Logistic Regression to be most effective and got accuracy of 95%, 77% and 80% respectively for disease prediction.

Quirina C.B.S. Thio et. al. [4] implemented five machine learning models for survival prediction and concluded Random Forest to most effective closely followed by support vector machine.

Shinya Iwase et. al. [5] implemented multiple machine learning models to predict ICU Mortality. They identified lactate, lactate dehydrogenase and platelet count as key variables. And found Random Forest to be the best machine learning model with accuracy, specificity and brier score of 94.5%, 87.5% and 0.028 respectively.

Jian Hu et. al. [6] used radial basis function artificial neural network model to predict mortality of acute respiratory distress syndrome (ARDS). They found LDH, organ failure, SP-D and PaO2/FiO2 to be the most important independent variables. They achieved sensitivity, specificity and accuracy as 83.6%, 88.5% and 82.5%, respectively for mortality prediction for ARDS ANN.

Min Chen et. al. [7] proposed a new convolutional neural network based multimodal disease risk prediction using structured and unstructured data from hospitals for regional chronic disease of cerebral infraction. Compared to several typical prediction algorithms, the prediction accuracy of their proposed algorithm reached 94.8% with convergence speed which was faster than that of CNN-based unimodal disease risk prediction.

Ameena N. Husain et. al. [8] developed an EMR pipeline to automate nSoFA calculation for more granular analysis of score performance and to identify optimal alerting thresholds.

Luís A. Vale-Silva et. al. [9] proposed MultiSurv, a multimodal deep learning method for long-term pan-cancer survival prediction which used dedicated sub models to establish feature representation of different data modalities.

Yannick Suter et. al. [10] evaluated 3D-convolutional neural networks (CNNs) and classical regression methods with hand-crafted features for survival time regression of patients with high-grade brain tumour. The tested CNNs for regression showed promising but unstable results. The best performing deep learning approach reached an accuracy of 51.5% on held-out samples of the training set. All tested deep learning experiments were outperformed by a Support Vector Classifier (SVC) using 30 radiomic features.

Zhenchao Sun et. al. [11] proposed an innovative model based on Graph Neural Networks (GNNs) for disease prediction, which utilizes external knowledge bases to augment the insufficient EMR data, and learns highly representative node embeddings for patients, diseases and symptoms from the medical concept graph and patient record graph respectively constructed from the medical knowledge base and EMRs. Their model offered an intuitive yet accurate solution to disease prediction, tackling the data scarcity problem and the hardship in diagnosing rare diseases at the same time.

Alghatani K et. al. [12] performed predictive modelling with minimal features while maintaining reasonable performance using the quantiles approach. The best accuracy achieved in the mortality model was approximately 89% using the random forest algorithm. The highest accuracy achieved in the length of stay model, based on the population median ICU stay (2.64 days), was approximately 65% using the random forest algorithm.

1. **METHODOLOGY**

**3.1 Data Cleaning**

The first step was data cleaning, which involved removing irrelevant data, dealing with missing values and outliers and correcting any inconsistencies in the dataset. Data cleaning ensured that the dataset is reliable and of good quality, thereby enabling accurate prediction of patient disease and mortality.

**3.2 Data Preprocessing**

The next step was data preprocessing, which involved transforming the data into a format suitable for analysis. This included scaling, normalization, and Label encoding of categorical variables.

**3.3 Feature Selection and Extraction**

Feature selection and extraction is the process of selecting the most relevant features for use in the prediction model. This was done using pipeline to drop constant, correlated and constant features. Some features were also dropped manually with intelligent guesswork. The aim of feature selection and extraction was to reduce the dimensionality of the dataset while retaining the most informative features.

**3.4 Model Selection and Training**

Model selection and training involved selecting appropriate machine learning and artificial intelligence models for predicting patient outcomes. Multiple models were used for prediction including logistic regression, decision tree, random forest, support vector machine, and artificial neural network for additional comparative studies. Model training involved fitting the selected models to the preprocessed data using appropriate algorithms. During training, the models were adjusted to optimize their performance on the training data.

**3.5 Performance Evaluation**

After training, we evaluate the performance of each model on a separate test set. We measure the performance of each model based on several metrics, including accuracy, precision, recall F1-score and brier score. We use these metrics to compare the effectiveness of each model and identify their strengths and weaknesses.

**3.6 Prediction Based on Performance**

The best-performing model was then selected and used to predict patient mortality and disease progression. The model is applied to new data, and predictions are made based on the performance metrics determined during evaluation. The predictions are used to provide recommendations for clinicians to improve patient outcomes. The prediction results are also compared with other models to demonstrate the effectiveness of the methodology.

1. **PREDICTION ALGORITHMS**
   1. **Decision Tree Classifier**

The decision tree classifier (DTC) is one of the possible approaches to multistage decision making; able look-up rules, decision table conversion to optimal decision trees and sequential approaches are others. The basic idea involved in any multistage approach is to break up a complex decision into a union of several simpler decisions, hoping the final solution obtained this way would resemble the intended desired solution. [13].

* 1. **Logistic Regression**

Logistic regression models are used to study effects of predictor variables

on categorical outcomes.

* 1. **Random Forest Classifier**

The random forest was first proposed by Leo Breiman from the University of California in 2001. It is composed of many basic classifiers (decision tree) which are completely independent from each other. Input a test sample to the new classifier and the class label of this sample can be decided based on the voting results from each single classification. The random operation is introduced in the build process, including the selection of samples subset and feature subset, to guarantee the independence of each decision tree, improve classification accuracy and gain better generalization ability. [14]

* 1. **KNN**

Classification based on the k-nearest neighbour algorithm differs from the other methods considered here, as this algorithm uses the data directly for classification, without building a model first. As such, no details of model construction need to be considered, and the only adjustable parameter in the model is k, the number of nearest neighbours to include in the estimate of class membership: the value of P(y|x) is calculated simply as the ratio of members of class y among the k nearest neighbours of x. By varying k, the model can be made more or less flexible (small or large values of k, respectively).

The advantage that k-nearest neighbours have over other algorithms is the fact that the neighbours can provide an explanation for the classification result; this case-based explanation can provide an advantage in areas where black-box models are inadequate.

The major drawback of k-nearest neighbours lies in the calculation of the case neighbourhood: for this, one needs to define a metric that measures the distance between data items. In most application areas, it is not clear how to, other than by trial and error, define a metric in such a way that the relative (but unknown!) importance of data components is reflected in the metric. [15]

* 1. **Support Vector Machine**

These models are algorithmic implementations of ideas from statistical learning theory, which concerns itself with the problem of building consistent estimators from data: how can the performance of a model on an unknown data set be estimated, given only characteristics of the model, and performance on a training set?

Algorithmically, support vector machines build optimal separating boundaries between data sets by solving a constrained quadratic optimization problem. By using different kernel functions, varying degrees of nonlinearity and flexibility can be included in the model. Because they can be derived from advanced statistical ideas, and bounds on the generalization error can be calculated for them, support vector machines have received considerable research interest over the past years. Performances on par with or exceeding that of other machine learning algorithms have been reported in the medical literature.

The disadvantage of support vector machines is that the classification result is purely dichotomous, and no probability of class membership is given. [15]

* 1. **Naive Bayes**

Naive Bayes is a collection of classification algorithms which are based on Bayes Theorem. Naive Bayes classifier gives us an excellent result when one uses it for text data analysis. Such as Natural Language Processing. Naive Bayes algorithm gives us a probability analysing the data set we have given. Naïve Bayes classifier is used as a probabilistic classifier. To perform the classifier, it uses the

concepts of mixture models. A mixture model is capable of establishing the probability of the component that it consists of Bayes theorem to perform as a probabilistic classifier. Another name that a naïve Bayes is known as simple Bayes or independence Bayes [16]. The probability P is defined as follows:

**P(C|X) = (P(X|C) \* P(C)) / P(X)**

P(C|X) is the posterior probability of class C given the observed features X.

P(X|C) is the likelihood of observing the features X given class C.

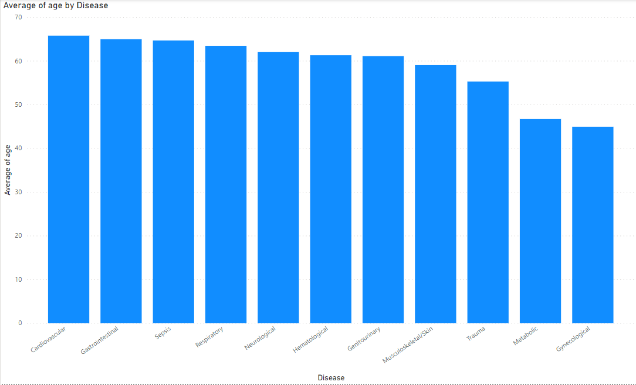
P(C) is the prior probability of class C.

P(X) is the probability of observing the features X.

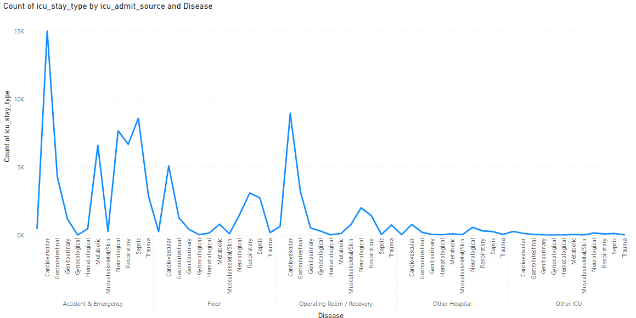
**RESULTS**

After performing the experiments and statistical analysis, we obtained the following results:

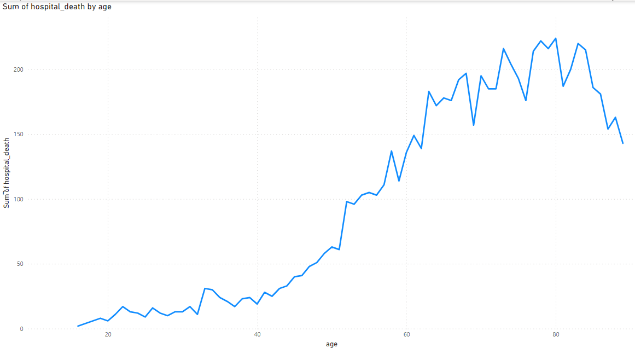
After thorough data analysis, it was observed that individuals in the late 50s and early 60s tend to have a higher likelihood of contracting various diseases. However, it is important to note that metabolic and gynaecological diseases show a different pattern, with a higher incidence among individuals in their mid-40s.



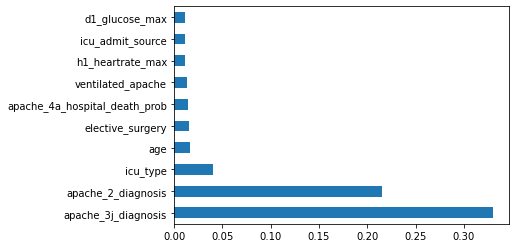
Upon conducting an extensive analysis, it was discovered that among the various medical conditions leading to ICU admissions, cardiovascular diseases exhibited the highest prevalence. This was closely followed by sepsis, respiratory ailments, and metabolic disorders, indicating their significant impact on ICU occupancy. Additionally, it is noteworthy that accidents and emergencies emerged as the primary source of ICU admissions, consistently contributing to a substantial number of patients requiring urgent critical care and specialized medical attention.



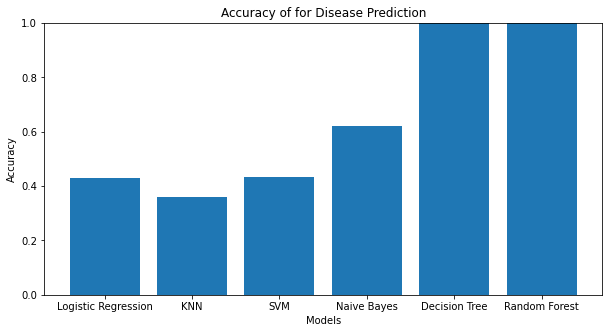
Through a comprehensive analysis, it has been determined that individuals aged 60 and above made the most significant contributions to the overall mortality rate. Among the various causes of death, cardiovascular diseases emerged as the most fatal, primarily affecting patients in the aforementioned age group. In contrast, trauma exhibited a more uniform impact, affecting patients across all age ranges without a distinct age-related pattern.



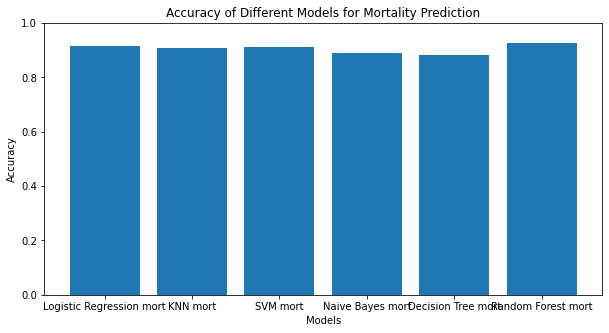
The most important features for prediction calculated through ExtraTreesClassifier came out as diagnosis, icu type and age.



Decision tree classifier and random forest classifier turned out to be most effective for disease prediction among all six machine learning models with accuracy of about 99% accuracy.



In case of mortality prediction using EMRs almost all the machine learning models performed equally with Random Forest classifier performing the best with accuracy of 92.6%.



**CONCLUSION**

In conclusion, the use of machine learning and artificial intelligence models for patient mortality and disease prediction from electronic medical records (EMRs) has the potential to significantly improve patient outcomes.

The results of this study demonstrate the effectiveness of machine learning and artificial intelligence models in accurately predicting patient mortality and disease progression. These predictions can provide valuable insights for clinicians and help them develop personalized treatment plans for their patients.

With continued research and development, machine learning and artificial intelligence can become even more valuable in improving patient care and outcomes. Overall, the methodology presented in this research paper has significant implications for the future of healthcare, as it can help clinicians provide more personalized and effective care for their patients, ultimately leading to improved patient outcomes.

# **References**

|  |  |
| --- | --- |
| [1] | P. G. P. P. E. M. Dhiraj Dahiwade, "Designing Disease Prediction Model Using Machine Learning Approach," in *3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019. |
| [2] | L. &. D. C. &. V. P. Sharmila, "Disease Classification Using Machine Learning Algorithms - A Comparative Study," *International Journal of Pure and Applied Mathematics. ,* vol. 114, pp. 1-10, 2017. |
| [3] | P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," in *4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2018. |
| [4] | K. A. B. B. Thio QCBS, "Development and Internal Validation of Machine Learning Algorithms for Preoperative Survival Prediction of Extremity Metastatic Disease," *Clinical Orthopaedics and Related Research,* vol. 478(2), pp. 322-333, 2020. |
| [5] | T.-a. N. T. S. T. O. T. S. N. T. J. Y. Y. Y. &. E. K. Shinya Iwase, "Prediction algorithm for ICU mortality and length of stay using machine learning.," *Scientific Reports,* vol. 12, 2022. |
| [6] | J. F. Y. &. L. W. Hu, "Predicting the mortality risk of acute respiratory distress syndrome: radial basis function artificial neural network model versus logistic regression model.," *J Clin Monit Comput,* vol. 36, p. 839–848, 2022. |
| [7] | Y. H. K. H. L. W. a. L. W. M. Chen, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," *IEEE Access,* vol. 5, pp. 8869-8879, 2017. |
| [8] | A. E. E. &. V. Z. Husain, "Use of an electronic medical record to optimize a neonatal sepsis score for mortality prediction.," *Journal of perinatology : official journal of the California Perinatal Association,* vol. 10, 2022. |
| [9] | L. R. K. Vale-Silva, "Long-term cancer survival prediction using multimodal deep learning," *Scientific Reports,* vol. 11, 2021. |
| [10] | Y. J. A. R. M. K. U. H. E. W. R. R. &. M. Suter, "Deep Learning Versus Classical Regression for Brain Tumor Patient Survival Prediction," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer International Publishing, 2019, pp. 429-440. |
| [11] | Y. H. C. H. C. T. C. L. Y. F. Sun Z, "Disease Prediction via Graph Neural Networks.," *IEEE Journal of Biomedical and Health Informatics,* vol. 25(3), pp. 818-826, 2021. |
| [12] | A. N. R. A. S.-N. A. Alghatani K, "Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation.," *JMIR Med Inform,* vol. 9(5), 2021. |
| [13] | S. R. S. a. D. Landgrebe, "A survey of decision tree classifier methodology," *Transactions on Systems, Man, and Cybernetics,* vol. 21, no. 3, pp. 660-674, 1991. |
| [14] | A. K. R. P. V. Parmar, "A Review on Random Forest: An Ensemble Classifier," in *Springer*, 2018. |
| [15] | L. O.-M. Stephan Dreiseitl, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of Biomedical Informatics,* vol. 35, no. 5-6, pp. 352-359, 2002. |
| [16] | A. K. a. A. K. M. M. A. M. Rahat, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in *International Conference System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India, 2019. |