A Patient Care Predictive Model using Logistic Regression

Harkesh J. Patel, Jatinderkumar R. Saini*
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed University), Pune, India

***Abstract*—Medical treatments and operations in hospitals are divided into in-patient and out-patient procedures. It is critical for patients to know and understand the differentiation between these two forms of treatment since it will affect the time of a patient's stay in a hospital or a medical institution as well as the cost of a treatment. In today's era of information, a person's talents and expertise may be put to good use by automating activities wherever possible. A medical service will be termed inpatient care if a doctor issues an order and the patient is admitted to the hospital on that order whereas a patient seeking outpatient care do not need to spend the night in a hospital. Choosing between in-patient and out-patient care is usually a matter of how involved the doctor wants to be with the patient's treatment. With the aid of numerous data points regarding the patients, their illnesses, and lab tests, our main objective is to develop a system as part of the hospital automation system that predicts and estimates whether the patient should be given an in-patient care or an out-patient care. The main idea of the paper is to understand and develop a logistic regression model to predict whether a patient needs to be treated as an in-patient or an out-patient depending on the results of laboratory tests. Furthermore, this study also focuses on how logistic regression performs for this dataset. In addition, research on how logistic regression performs for this dataset was also not done. From the study, the results show that logistic regression gives an accuracy of 75%, F1-score of 73%, precision of 74% and recall of 74%.**

***Keywords*—Health-care; inpatient care; logistic regression; machine learning model; outpatient care; stacking classifier**

I. Introduction

Inpatient and outpatient procedures can be used to classify medical treatments and operations. A patient should be aware of this distinction since it affects how long he or she stays in a hospital and how much the operation costs. The difference between an in-patient and an out-patient treatment is the duration of time a patient must spend time in a medical institution where the operation or treatment is performed. In-patient treatment compulsorily requires an overnight stay in the hospital or the medical institution.

Patients must spend time or stay at least for one night in the medical facility that is provided to them where their operation or treatment was performed, typically a hospital. During this period, kids are normally under the care of a nurse or a doctor.

Patients seeking out-patient care need not have to spend the night in a hospital. These patients are allowed to depart the hospital once their treatment is completed. In certain situations, they must wait while the anesthetic wears off or to ensure that there are no problems. Patients are not required to spend the night under supervision if there are no significant problems. In addition, the out-patient treatment is much more cost-effective in comparison to in-patient treatment [12].

To categorise patients' medical requirements into an in-patient or an out-patient treatment, doctors rely heavily on the results of lab tests. This time-consuming approach requires doctors to exert considerable effort in order to determine if the patient is required to be in the hospital and closely watched or not. Also, the patient's life may be in risk if the wrong decision is made [1]. Clinical machine learning research is largely restricted to proof-of-concept studies. Machine learning applications in clinical medicine are currently hindered by a number of obstacles. A major shift in medical practise may result from overcoming obstacles to potential deployment, with the use of specialised technologies helping the healthcare team provide better, more customised patient care [2].

Machine learning algorithms, it is widely accepted, find and extract information based on the data available. Yet a vast quantity of information is available in machine-readable format, ready to be incorporated into machine learning algorithms and models [4]. For this study, Logistic Regression, a supervised machine learning algorithm, is used. The main objective of this paper is to see how logistic regression works on such a problem. Furthermore, we will check the performance metrics, precision, recall and accuracy to get an idea of how useful logistic regression is for such problems.

II. Literature Review

Melhem et al. [1] built four models depending on the patient's circumstances and lab test results: support vector machine model, decision tree model, random forest model and k-nearest neighbours model. The major aim of their study was to make use of ML algorithms to categorize the patient treatment as an in-patient or out-patient, in order to lessen the time and effort expended by the healthcare experts, which reflects the kind of services provided to the patient. Furthermore, this research assists in the reduction of human errors, which can result in hazards to the patient's life as well as an increase in the overall bill amount. The best model out of four was picked based on its accuracy, sensitivity, specificity, and precision scores, as well as its low false-negative & false-positive rates. To construct and evaluate these models, the EHR dataset was utilised, which comprises of patients' laboratory test results from a private hospital that is in Indonesia. The outcomes of their study say that random forest algorithm had the best accuracy (77 %), precision-rate (72%) & sensitivity (65%) as well as the model had the lowest false-

*Corresponding Author

negative rate (35%) and almost the lowest false-positive rate (16%).

Ben-Israel et al. [2] conducted a review in keeping with the prisma criteria and concentrating on human studies which utilised machine learning to directly treat a hard-headed scenario. The studies were performed between 1st Jan, 2000 & 1st May, 2018 and offered data on the performance of the used machine learning technique. Reviewers looked over 1909 distinct publications and found 378 retrospective papers and 8 prospective ones that met eligibility requirements. 61% of papers published in the past four years were retrospective. Few articles met our inclusion criteria, with just 2% of them being prospection articles. When it comes to clinical medicine, a majority of the literature is retrospective and focuses on proof-of-concept ways to improving patient care. A major transformation in medical practise will be enabled by recognizing the key translational hurdles, including instantaneous access to hard-headed data, data reliability, medical practitioner approval of "black box" generated findings and performance evaluation.

Beaulieu-Jones et al. [3] have talked about the applications of ML in health-care which is increasing quickly and might have a major effect on the profession. Using ML for health-care research, the study was aimed to provide quantitative and qualitative assessments of the current status. In order to assess the present status of research in ML for health-care, including areas of methodological and medical focus and limits, as well as areas that are underexplored, they analysed contributions. Results showed that the clinical collaborators were involved in 58 (34.9%) of the 166 accepted entries, and in 83 (50.0%) of the submissions that focused on clinical practise. On average, (97 datasets) 58% of the data sets utilised were publicly available or needed registration. (70 articles (42.2%)) of them were in clinical practise, with brain & mental health (25(15.1%)), cancer (21 (12.7%)) and cardiovascular (19 (11.4%)) being the most prevalent specialities. Data that is well-annotated and freely accessible is critical to the development of translational implementations in ML for health-care research, according to current trends.

Radovanović et al. [4] proposed an approach to logistic regression that incorporates domain information in the form of ontologies/hierarchies via layered generalisation. Because ontology/hierarchy relations are stacked, they may be combined to create higher, abstract ideas. In this case, they were able to tackle the problem of unexpected 30-day hospital readmissions. The proposed framework outperforms ridge, lasso and tree lasso logistic regression in terms of accuracy. This framework increases AUC by up to 9.5% for children and up to 4% for severely over-weight patients. Also, it increases the AUPRC up to 5.7% for children and up to 2.6% for ghastly over-weight patients, the researchers found.

Mu-Yen Chen [5], when using Decision Tree (DT) classification, the accuracy decreased the more PCA was applied, according to this paper's findings. According to his findings in the study of financial hardship, the accuracy of their DT classification technique increased as time passed, with an accuracy rate of 97.01% for two seasons previous to financial difficulty. They found that PCA increases the error while attempting to identify firms in a financial crisis as "normal" enterprises, and that DT classification has a higher short-term prediction accuracy than the Logistic Regression (LR) classification technique (less one year). Instead, the LR method improves long-term prediction accuracy (above one and half year). A short-term financial distress prediction model using AI, rather than standard statistical methods, is proposed in this study as a possible alternative to traditional statistical methods.

El-Rashidy et al. [6] have come up with a new way to forecast ICU patient death using stacking ensemble. Compared to the literature study, their method is more accurate and intuitive from a medical point of view in collaboration with an ICU domain specialist, data were produced and features selected. On the basis of the expert's judgments, six categories of data were created. When it came to the prediction procedure, each modality was assigned a distinct classifier depending on its performance. Our classifiers included linear discriminant analysis, decision tree algorithm, multilayer perceptron, k-nearest neighbour (knn) and logistic regression, among others. A stacking ensemble classifier was then created and tuned using the five classifier decisions. The system was validated with the help of a benchmark dataset of over ten thousand patients from MIMIC III. Patients' time series data of varied durations was used to undertake extensive studies in order to predict death. The first six, twelve, and twenty-four hours of a patient's initial stay were tested. On the basis of the results, their model surpassed the current techniques in terms of accuracy i.e. 94.4%, f1-score i.e. 93.7%, precision i.e. 96.4%, recall i.e. 91.1% & ROC curve i.e. 93.3%. As a result of these findings, it's clear that their technique of predicting ICU mortality works.

Polikar [7] studied situations where ensemble-based systems are superior to single-classifier systems, techniques for producing separate parts of ensemble systems and methods for combining separate classifiers. Many ensemble-based algorithms like bagging and boosting have been discussed as well as generalisation and hierarchical mixtures of experts. They have also discussed typically used combination rules such as algebraic blend of outputs, voting-based approaches and behaviour knowledge space as well as decision templates. A last look at future exploration prospects for ensemble systems was conducted. In addition, ensemble systems have showed significant promise in a variety of other fields like as feature selection and learning with lost features, confidence estimation and fault-correcting output codes. It has been demonstrated that ensemble-based systems generate better outcomes than expert systems for a broad range of implementations and circumstances. In their paper, they have talked about how to design, build, and use such systems.

Saini et al. [8] in their study seek to present the importance of artificial intelligence in magnetic learning algorithm and examines their function in different sectors of health, such as bioinformatics, cancer gene identification, epileptic seizure, brain computer interface. It also examines the medical imaging of illnesses, including diabetic retinopathy, gastro-intestinal disease, and tumour via extensive learning. Finally, this essay highlights the real barriers to AI approaches which need to be addressed. They examined the reason why ML was used in healthcare in this work. The main category of ML, is also discussed. They concentrated on deep learning, its architecture and explore various health data analysing and examining deep learning. However, ML technologies draw considerable attention in the field of medical research. There are still difficulties with real-time implementation. Regulations are one such difficulty. Recent rules lack safety, evaluation and efficiency criteria for the ML system. The US FDA provides advice for the evaluation of ML systems to preserve security and efficiency in order to solve this challenge. The existing health care environment does not encourage the exchange of information on the system. It is also a limitation. The ML training was therefore compromised before implementation. In many nations, the Health Care Revolution encourages data exchange.

Kirasich et al. [13] addresses the challenge of model selection by assessing the overall classification performance for datasets with diverse underlying structures between the random forest and logistic regression by increasing the variance in the explanatory as well as the noise variables, number of explanatory variables, noise variables and observations. They created a model evaluation tool which can simulate classification models for such data and performance indicators as real positive rates, false positives and accuracy in certain circumstances. They observed that logistic regression has continuously exercised a better overall precision than random forests by increasing the variance not only in the explanatory but also the noise factors. The true positive-rate for random forest algorithm was, however, greater than the logistic regression algorithm & the data set with rising noise factors showed higher false-positive rates. Each and every case study included thousand simulations and the model executions in it consistently demonstrated that the false positive-rate for random forest with hundred trees was scientifically different from logistic regression. Under varied simulated dataset circumstances, logistic regression algorithm & random forest algorithm produced variable corresponding classification scores in all four situations.

Maroof [14] says that based on the continuous predictor, logistic regression seeks to classify or predict a discrete, categorical variable from among continuous or discrete predictors. Clinical neuropsychology's preference for using this paradigm in research is connected to the discipline's fundamental structure, which includes the use of scientific terminology to explain cognition and behaviour, as well as the compartmentalization of syndromes into diagnostic entities.

Goldarag et al. [15] developed, tested and compared forest fire risk prediction models based on logistic regression and neural networks. The findings show that the neural network model is more accurate at categorising fire points than logistic regression, which is sensitive to fire point samples. The percentage of fire and non-fire samples must be matched to obtain high accuracy in logistic regression. A neural network with two hidden layers, twenty-eight neurons, and a logarithmic-sigmoid transfer function in both hidden layers was also tested and the best architecture was found to be a neural network with two hidden layers, 28 neurons, and a logarithmic-sigmoid transfer function in both hidden layers.

Pepe & Thompson [16] developed strategies for maximising the accuracy of routinely used diagnostic measures by discovering linear combinations of markers. The approaches were non-distributional, appeared to have strong statistical features, and could account for heterogeneity defined by variables.

Sadikin and Mujiono [17] created an electronic health record predicting dataset obtained from a private hospital in Indonesia. It comprises the findings of the patient's laboratory tests, which are used to determine the next patient treatment, whether the patient is in or out of the hospital.

In order to determine the allocation of staff care in community-acquired pneumonia, España et al. [18] created a new prediction algorithm based on the 5 risk classifications described by the Pneumonia Severity Index. There was no question about the decision to hospitalise low risk (I-III) classes, when one or more of the following was evident, namely: tension in arterial oxygen < 8.0 kPa (60 mm Hg), shock, coexisting decompensating diseases, plural effusion, unable to maintain oral intake, a social problem and a lack of reaction to prior empirical antibiotic therapy. The findings are presented in a number of 616 patients after 18 months following application of this new prediction criteria. In 221 patients treated as ambulatory patients, the death rate was 0.5% vs 8.9% in 395 patients treated as hospitals. Of the 178 low risk individuals treated as hospital patients, 106 were given the specific extra requirements for hospitalisation, while the other 72 evidently did not justify the decision to admit hospitalisation under the predictive criterion. These 72 patients had better results than high-risk patient and low-risk patient who fulfilled the extra particular criteria for admitting to hospital (substantially shorter admission, antibiotic days, death and complex course) their results were better. Therefore, rigorous adherence to the new prediction criteria might have prevented admission in these low-risk individuals. Another significant finding was that not all patients admitted to the hospital had been identified in the Pneumonia Severity Index alone.

Blais et al. [19], in their study identified factors related with length of stay (LOS) and included measurement of these variables into their normal preadmission evaluation. A retrospective research of 80 discharged patients looked at the relationship between LOS and 25 factors representing a combination of patient/demographic characteristics, disease variables, and therapy variables. According to multivariate analysis, ten factors independently accounted for 62% of the variance in LOS. The information utilised was largely gathered during the pre-admission screening. In the prospective research, the factors' predictive ability decreased. However, fewer individual factors were substantially related to LOS; the total of the variables' scores predicted 17% of the LOS variation. The findings showed that significant criteria for predicting LOS are accessible at the time of admission, and these variables may be systematically examined and incorporated into clinical decision making.

Cuffel et al. [20] examined the correctness of different models for projecting a rehospitalization in a maintained mental health organisation, as well as the efficacy of various care or treatment management methods for improving out-patient treatment follow-up. In a randomised controlled trial, patients that had in-patient mental health or substance use admissions were assigned to one out of 3 types of treatment supervision based on the level of involvement of treatment executives in the release planning & post-release outreach i.e. usual (N=31), enhanced (N=94) or intensive (N=74). Here, the classes that were formed were compared to each other and to a cohort hospitalised the year before the research and given usual treatment (N=192) to see if there were any changes in time to out-patient check out, quantity of post-release care and rehospitalization at 30, 60 & 180 days. There weren't any differences between the classes found. The larger part of number of patients i.e. 69% got out-patient treatment within 30 days of being released. The number of hard-headed and socio-demographic risk variables reported by care executives was associated to the probability of rehospitalization at 60 and 180 days, according to logistic regression prediction models. Patients who were approved to receive intermediate treatment i.e. partial hospitalisation and those who did not attend the intermediate treatment, if it was authorised were more suitable to be re-hospitalized at 30, 60 & 180 days than other patients. With increasingly extensive release planning and outreach, outpatient follow-up following mental hospitalisation did not improve. Improved prediction of re-hospitalization risk may improve possibilities to deliver intensive treatments to difficult-to-engage patients.

III. Methodology

Most of the data science implementations depend heavily on ML models. Other expert knowledge exists outside of the given data, which may theoretically assist the ML algorithms better recognize the conditions and circumstances of the data that is provided [4].

Machine Learning employs three types of learning: supervised, unsupervised and semi-supervised learning.

Supervised learning is a type of training in which we educate or train the machine using well-labelled data, i.e. data that already has the correct answer. Following that, the machine is given a fresh collection of instances, i.e. the test data, so that the supervised learning algorithm may analyse the training data and generate a proper result from labelled data. Further, supervised learning approaches include regression and classification, which are further classified. Regression is a helpful statistical prediction approach that aims to establish a meaningful link between dependent and independent variables by attempting to find correlations between them. To forecast a continuous output, the regression technique is employed in machine learning (ML). The predicted result is a real number. The output of classification is discrete, but the output of categorization is continuous [8].

Unsupervised learning, on the other hand, is the training of a computer utilising input that has not been categorised or labelled and enabling the algorithm to operate on that information without supervision. Clustering and principal component analysis (PCA) are two of the most important techniques in unsupervised learning. PCA is typically used to reduce the size of an object. With numerous dimensions, PCA reduces the data to a few principle component directions without losing much of the data. PCA is often used before clustering to minimise the number of dimensions of the data before it is clustered. Instead of using output information, the clustering approach is used to create a collection of variables that exhibit similarities or commonalities. As a result of these algorithms, the cluster labels for the variable with the highest degree of similarity within and between the clusters are generated and displayed on a graph [8].

Semi-supervised learning, on the other hand, includes function estimation on both labelled and unlabelled data. This method is driven by the fact that labelled data is frequently expensive to create, but unlabelled data is cheap.

Here, we have used Logistic Regression as it a binary classification task i.e. 0 or 1. A discrete, categorical variable is classified or predicted using logistic regression using continuous or discrete predictor, such as yes/no depending on the continuous predictor [14]. In order to achieve high accuracy in logistic regression, the percentage of in-patient and out-patient care samples must be balanced [15]. It produces a linear score that clearly distinguishes between two outcomes [16]. Moreover, the research on how logistic regression model performs on such scenario was also not done. General approaches such as ensemble learning can be used to improve the accuracy of prediction or classification models such as decision trees and artificial neural networks [11]. Hence, Stacking Classifier is used for improving the accuracy of prediction.

A. *Dataset and Experimental Discussion*

The dataset comprises predictions from an Electronic Health Record gathered from a private hospital which in Indonesia. It comprises the laboratory test results of different patients, which are used to decide the next patient's treatment, whether in or out of the hospital. The dataset contains 4412 rows and 11 columns. The total number of features are 10, where number of numerical features are 9 and number of categorical feature is 1 [17].

B. *Attribute Information*

- HAEMATOCRIT: It is the patient's laboratory test result of haematocrit.
- HAEMOGLOBINS: It is the patient's laboratory test result of haemoglobins.
- ERYTHROCYTE: It is the patient's laboratory test result of erythrocyte.
- LEUCOCYTE: It is the patient's laboratory test result of leucocyte.
- THROMBOCYTE: It is the patient's laboratory test result of thrombocyte.
- MCH: It is the patient's laboratory test result of MCH.
- MCHC: It is the patient's laboratory test result of MCHC.
- MCV: It is the patient's laboratory test result of MCV.
- AGE: It is the patient's age.

- SEX: It is the patient's gender.
- SOURCE: Target i.e. Binary: in-patient/out-patient – 0/1.

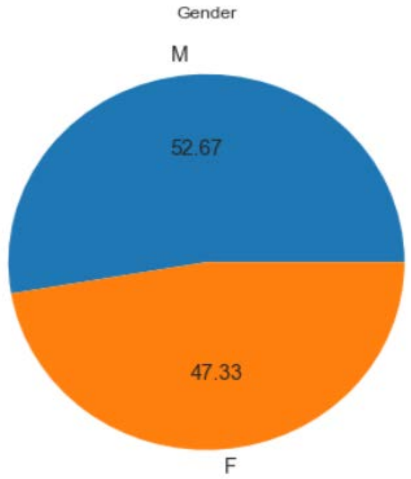

Fig. 1. Percentage of Males (M) and Females (F) in the Dataset.

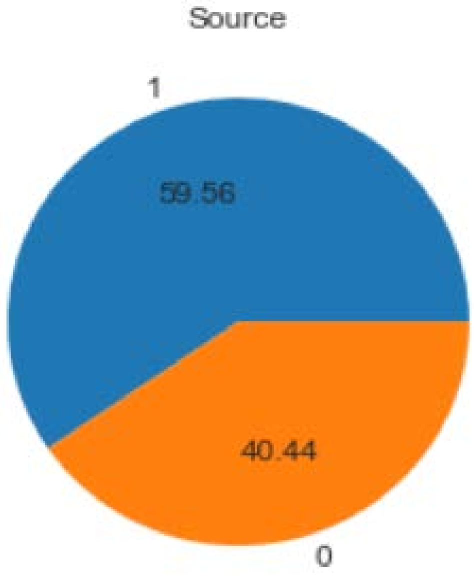

Fig. 2. Percentage of Inpatient (0) and Outpatient (1) in the Dataset.

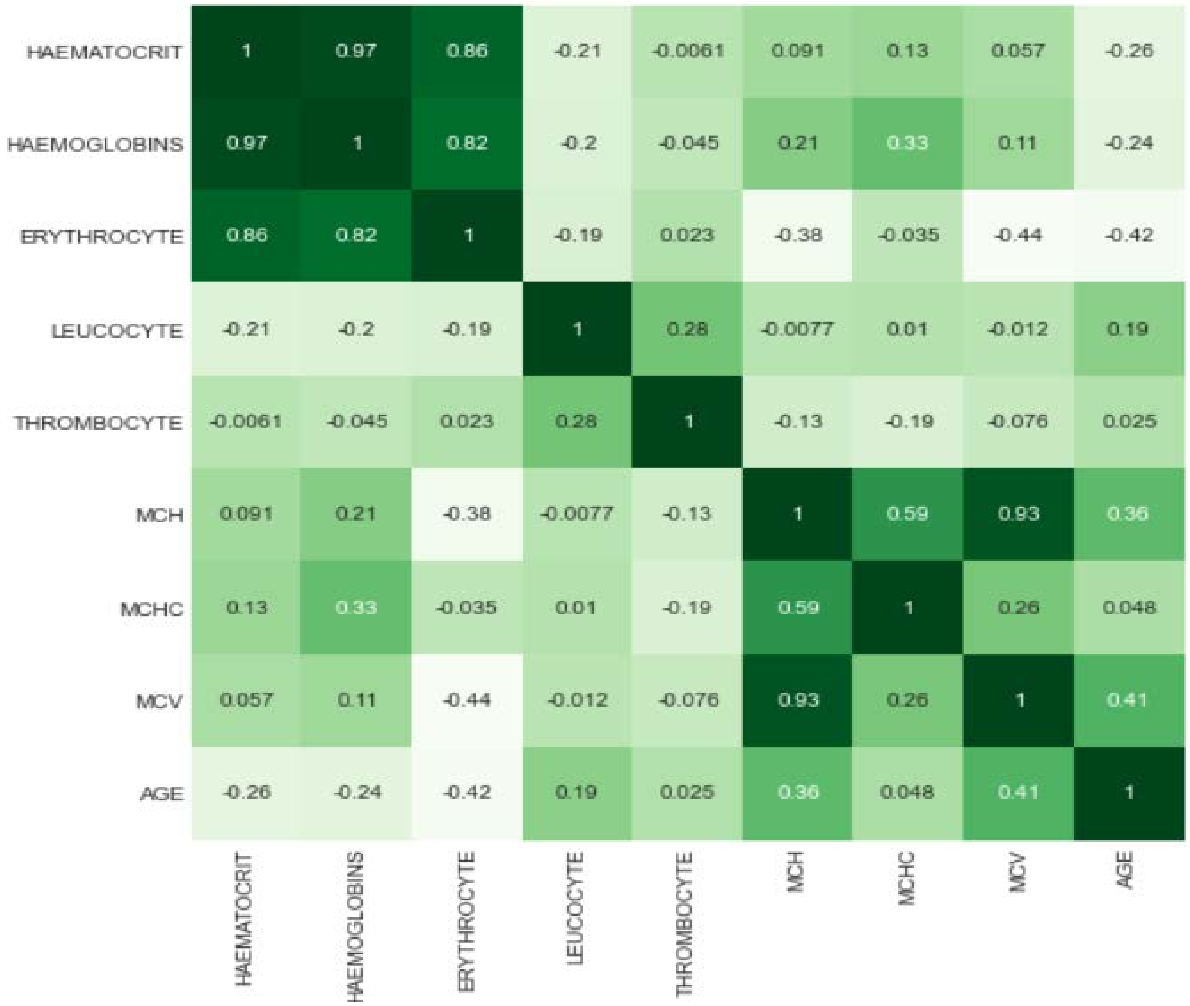

Fig. 3. Pearson's Correlation of Features with respect to each other.

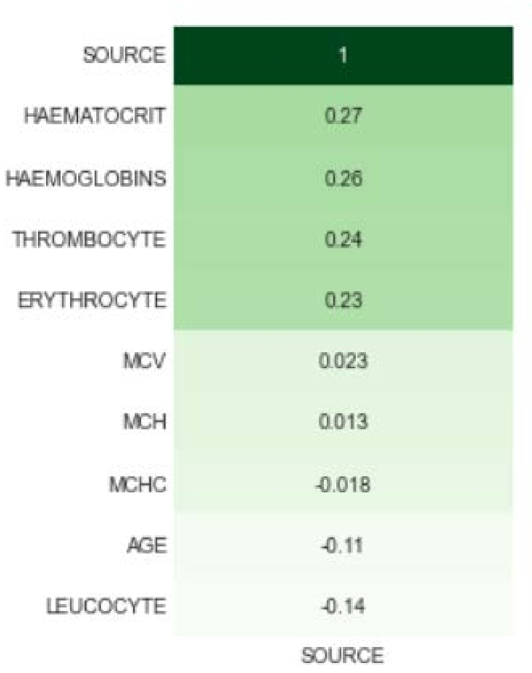

Fig. 4. Pearson’s Correlation of Features with respect to Target.

TABLE I. FIRST FEW RECORDS OF THE DATASET

HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	AGE	SEX	SOURCE
35.1	11.8	4.65	6.3	310	25.4	33.6	75.5	1	F	1
43.5	14.8	5.39	12.7	334	27.5	34.0	80.7	1	F	1
33.5	11.3	4.74	13.2	305	23.8	33.7	70.7	1	F	1

TABLE II. PERCENTAGE OF MISSING VALUES IN EACH FEATURE

Attribute	Missing %
HAEMATOCRIT	0.0
HAEMOGLOBINS	0.0
ERYTHROCYTE	0.0
LEUCOCYTE	0.0
THROMBOCYTE	0.0
MCH	0.0
MCHC	0.0
MCV	0.0
AGE	0.0
SEX	0.0
SOURCE	0.0

The pie chart in Fig. 1 shows that there are 52.67% Males (M) and 47.33% Females (F) in the dataset. While the pie chart in Fig. 2 shows that 40.44% people were treated as inpatient (0) and 59.56% people were treated as outpatient (1) in the dataset. Fig. 3 shows the Pearson’s Correlation of features with respect to each other whereas Fig. 4 shows the Pearson's Correlation of features with respect to the target i.e. SOURCE.

In the dataset, the 80% data was taken as training data and 20% data was taken as test data. The Tables I, II and III show the first few records of the dataset, percentage of missing values in each column and number of unique values in each column respectively. During the feature engineering part, we replaced the labels of sex column with binary numbers i.e. F = 0 and M = 1. Thereafter, MinMaxScaler was used to scale the features to a range of [0, 1]. At the end, we removed the least correlated features i.e. [MCH, MCHC, MCV] from the dataset. After performing feature engineering, my dataset was ready for the next step i.e. training the model. The exploratory data analysis and feature engineering was performed in order to increase the accuracy and precision of the model. Efficient and effective feature selection techniques allow better generalization of predictive models and improved interpretability, which is a very important property for applications in health care [10]. Furthermore, Hyperparameter tuning was performed for obtaining best case scenario.

Solving issues with a decision boundary which extends outside of the space of the function that is implemented by the specified classifier model is exceedingly difficult for a single classifier to do it successfully. This non-linear boundary may be learned by combining ensemble classifiers in the right way. To avoid bad selection of a single classifier that cannot generalise performance, combine many classifiers and average their output to lower the chance of poor performance of the single classifier that is picked. As a result, the chance of making a bad choice is reduced as well [6]. Thus, we have used Stacking Classifier in order to combine the skills of the models on the regression problem to produce predictions that outperform any single model in the ensemble.

It has been demonstrated that ensemble-based systems generate better outcomes than single-expert systems for a wide range of applications and circumstances [7].

TABLE III. NUMBER OF UNIQUE VALUES IN EACH FEATURE

Attribute	Number of unique values
HAEMATOCRIT	326
HAEMOGLOBINS	128
ERYTHROCYTE	433
LEUCOCYTE	276
THROMBOCYTE	554
MCH	189
MCHC	105
MCV	406
AGE	95
SEX	2
SOURCE	2

IV. Result Analysis

After using Logistic Regression, we got the train accuracy as 70.9% and test accuracy as 71.5%. For increasing the accuracy, we performed hyperparameter tuning using RandomizedSearchCV and got best parameters for logistic regression i.e. {'penalty': 'none', 'max_iter': 300, 'fit_intercept': True, 'class_weight': {0: 1, 1: 1}, 'C': 0.01}.

After retraining the model, we got the train accuracy as 71.6% and test accuracy as 73%. The Table IV shows the classification report of the model.

Stacking is one of the most efficient methods for solving classification and regression issues. The concept of stacking is using the predictions of machine learning models from the previous level as the input variables in the following level's machine learning models [9].

And hence, we have used Stacking classifier. We got the train accuracy as 72% and test accuracy as 75%. It could be seen that there was a minor increase in the accuracy. Furthermore, I have used cross-validation but stacking classifier gave a better accuracy and so, we went for stacking classifier. The Table V shows the classification report of the model.

TABLE IV. Classification Report of the Logistic Regression Model

	Precision	**Recall**	**F1-score**	**Support**
0 (In-patient Care)	0.75	0.52	0.61	357
1 (Out-patient Care)	0.73	0.88	0.80	526
Accuracy			0.73	883
Macro Average	0.74	0.70	0.70	883
Weighted Average	0.74	0.73	0.72	883

TABLE V. Classification Report of Logistic Regression using Stacking Classifier

	Precision	**Recall**	**F1-score**	**Support**
0 (In-patient Care)	0.74	0.53	0.62	357
1 (Out-patient Care)	0.73	0.88	0.80	526
Accuracy			0.75	883
Macro Average	0.74	0.70	0.72	883
Weighted Average	0.74	0.73	0.73	883

The findings of Melhem et al. [1] reveal that out of four models i.e. Support Vector Machine (SVM) model, Decision Tree model, Random Forest model and K-Nearest Neighbors (KNN) model, Random Forest model had the best accuracy (77%), precision (72%) and sensitivity (65%). So comparing with that model, our model gave nearly a similar accuracy (75%) but gave better precision (74%). Moreover, Kirasich et al. [13] observed that logistic regression has continuously exercised a better overall precision than random forest model by increasing the variance in the explanatory factors as well as noise factors. Under varied simulated dataset circumstances, logistic regression model and random forest model had produced variable relative classification scores in all the four situations they observed.

V. Conclusion and Future Work

The results show that the logistic regression gives nearly 75% accuracy, 73% recall, 73% f1-score on the test data. It gives a decent result on the dataset. Furthermore, the main objective and idea behind the research was fulfilled.

Moreover, for logistic regression model, independent characteristics may be used to predict accurate probabilistic outcomes based on statistical analysis. The model may over-fit on the training set if the dataset has a high number of dimensions, and hence may not be able to predict correct outcomes on the test set if the dataset has a large number of dimensions. Sometimes this happens if a little amount of data is used to train the model, but the data has a large number of features. Regularization strategies should be explored for high-dimensional datasets in order to avoid over-fitting but this makes the model complex.

Using Stacking classifier, there was a slight increase in the accuracy. Moreover, using multiple machine learning algorithms with stacking classifier or using regularization techniques on logistic regression with stacking classifier on a huge dataset could be thought of.

References

[1] S. Melhem, A. Al-Aiad and M. S. Al-Ayyad, "Patient care classification using machine learning techniques," 2021 12th International Conference on Information and Communication Systems (ICICS), 2021, pp. 57-62, doi: 10.1109/ICICS52457.2021.9464582.

[2] David Ben-Israel, W. Bradley Jacobs, Steve Casha, Stefan Lang, Won Hyung A. Ryu, Madeleine de Lotbiniere-Bassett, David W. Cadotte, The impact of machine learning on patient care: A systematic review, Artificial Intelligence in Medicine, Volume 103, 2020, 101785, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2019.101785.

[3] Beaulieu-Jones B, Finlayson SG, Chivers C, et al. Trends and Focus of Machine Learning Applications for Health Research. JAMA Netw Open. 2019;2(10):e1914051. doi:10.1001/jamanetworkopen.2019.14051.

[4] Sandro Radovanović, Boris Delibašić, Miloš Jovanović, Milan Vukićević and Milija Suknović A Framework for Integrating Domain Knowledge in Logistic Regression with Application to Hospital Readmission Prediction, International Journal on Artificial Intelligence Tools , VOL. 28, NO. 06, https://doi.org/10.1142/S0218213019600066.

[5] Mu-Yen Chen, Predicting corporate financial distress based on integration of decision tree classification and logistic regression, Expert Systems with Applications, Volume 38, Issue 9, 2011, Pages 11261-11272, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2011.02.173.

[6] N. El-Rashidy, S. El-Sappagh, T. Abuhmed, S. Abdelrazek and H. M. El-Bakry, "Intensive Care Unit Mortality Prediction: An Improved Patient-Specific Stacking Ensemble Model," in IEEE Access, vol. 8, pp. 133541-133564, 2020, doi: 10.1109/ACCESS.2020.3010556.

[7] R. Polikar, "Ensemble based systems in decision making," in IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21-45, Third Quarter 2006, doi: 10.1109/MCAS.2006.1688199.

[8] Saini, Akanksha and Meitei, A J and Singh, Jitenkumar, Machine Learning in Healthcare: A Review (April 26, 2021). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021, Available at SSRN: https://ssrn.com/abstract=3834096 or http://dx.doi.org/10.2139/ssrn.3834096.

[9] B. Pavlyshenko, "Using Stacking Approaches for Machine Learning Models," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018, pp. 255-258, doi: 10.1109/DSMP.2018.8478522.

[10] S. Radovanovic, M. Vukicevic, A. Kovacevic, G. Stiglic, and Z. Obradovic, “Domain knowledge Based Hierarchical Feature Selection for 30-Day Hospital Readmission Prediction,” Lecture Notes in Computer Science, pp. 96–100, 2015.

[11] King, Michael A. Ensemble learning techniques for structured and unstructured data. Diss. Virginia Polytechnic Institute and State University, 2015.

[12] Justin Oh, Anahi Perlas, Johnny Lau, Rajiv Gandhi, Vincent W.S. Chan, Functional outcome and cost-effectiveness of outpatient vs inpatient care for complex hind-foot and ankle surgery. A retrospective cohort study, Journal of Clinical Anesthesia, Volume 35, 2016, Pages 20-25, ISSN 0952-8180, https://doi.org/10.1016/j.jclinane.2016.07.014.

[13] Kirasich, Kaitlin; Smith, Trace; and Sadler, Bivin (2018) "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," SMU Data Science Review: Vol. 1 : No. 3, Article 9.

[14] Maroof D.A. (2012) Binary Logistic Regression. In: Statistical Methods in Neuropsychology. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3417-7_8.

[15] Jafari Goldarag, Y., Mohammadzadeh, A. & Ardakani, A.S. Fire Risk Assessment Using Neural Network and Logistic Regression. J Indian Soc Remote Sens 44, 885–894 (2016). https://doi.org/10.1007/s12524-016-0557-6.

[16] Margaret Sullivan Pepe, Mary Lou Thompson, Combining diagnostic test results to increase accuracy , Biostatistics, Volume 1, Issue 2, June 2000, Pages 123–140, https://doi.org/10.1093/biostatistics/1.2.123.

[17] Sadikin, Mujiono (2020), “EHR Dataset for Patient Treatment Classification”, Mendeley Data, V1, doi: 10.17632/7kv3rctx7m.1.

[18] P.P. España, A. Capelastegui, J.M. Quintana, A. Soto, I. Gorordo, M. García-Urbaneja, A. Bilbao, European Respiratory Journal 2003 21: 695-701; DOI: 10.1183/09031936.03.00057302.

[19] Blais, M.A., Matthews, J., Lipkis-Orlando, R. et al. Predicting Length of Stay on an Acute Care Medical Psychiatric Inpatient Service. Adm Policy Ment Health 31, 15–29 (2003). https://doi.org/10.1023/A:1026044106172.

[20] Cuffel, Brian J., Martin Held, and William Goldman. "Predictive models and the effectiveness of strategies for improving outpatient follow-up under managed care." Psychiatric Services 53.11 (2002): 1438-1443.