IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, March 2025

Medical Insurance Claim Validation System

Dr. K. K. Tripathi[1], Aamod Dabholkar[2], Gaurav Dongre[3], Omkar Jagdale[4]
Faculty, Department of Computer Engineering[1]
Student, Department of Computer Engineering[2,3,4]
Shivajirao S Jondhale College of Engineering, Dombivli (E), Thane, Maharashtra, India

Abstract*: The Medical Insurance Claims Verification System (MICVS) is a comprehensive framework designed to improve the efficiency and accuracy of the medical health insurance claims verification manner. In a quick-paced health care environment, powerful insurance declare agreement performs an important function in ensuring economic stability for fitness care companies and supplying uninterrupted carrier to patients. MICVS makes use of advanced technologies together with synthetic intelligence, system learning and information analytics to automate and streamline claims approval processes even as assembly regulatory requirements. The number one goal of MICVS is to reduce the risk associated with false claims and wrong reimbursements, thereby protecting the financial integrity of insurance carriers and reducing pointless charges. By combining superior algorithms and predictive modeling strategies, the gadget evaluates claims against pre- organized criteria, identifying anomalies, inconsistencies and capacity fraud with high accuracy and pace. In addition, MICVS facilitates actual-time records analysis and choice making, enabling speedy reaction to rising tendencies and styles in claims submissions*

Keywords: Health insurance claim validations, Machine learning, healthcare, regression, big data analytics

I. INTRODUCTION

1.1 Overview

Insurance fraud is a claim made for getting improper money and not actual amount of money from insurance company or any other underwriter. Motor and insurance area unit two outstanding segments that have seen spurt in fraud. Frauds is classified from a supply or nature purpose of read. Sources is client, negotiator or internal with the latter two being a lot of essential from control framework purpose of reads.

Fraud is common and costly to the health insurance system. Fraud is deception or misrepresentation intended solely for gain. That's surprising because health insurance fraud is increasing every year. Data mining techniques are used to detect and prevent fraud. It includes some basic knowledge of the health care system and its fraudulent behaviour, as well as an analysis of the characteristics of health insurance data. Data mining is divided into two learning methods, supervised and unsupervised, for the detection of false claims.

The medical community says that validation structures play an important role in ensuring the validity and legitimacy of health claims made by using patients, healthcare providers, and other stakeholders in the healthcare industry. This structure runs the backbone of insurance operations, facilitating green processing and settlement of claims, as well as protecting against fraud, error and abuse.

1.2 Objectives

- Publishing authentic proof to recognize patterns and symptoms of fraud. Applying engineering strategies toeliminate defects and tune models for efficiency.
- Train and adjust learning models of instruments including logistic regression, decision trees, and neural networks to classify demands appropriately.
- Evaluate the model's performance using thorough metrics and practical techniques.
- Integrate improved versions into Victory's claim validator to automate the system and improve overall performance.
- Constantly monitor and upgrade benchmarks to keep pace with changing fraud and maintain real-time productivity.

DOI: 10.48175/IJARSCT-23725

II. LITERATURE REVIEW

TABLE I

Sr.No	Survey Title	Author	Year	Paper Summary/Abstract
1	A Framework for Internal Fraud Risk Reduction at IT Integrating Business Processes: The IFR² Framework	Mieke Jans, Nadine Lybaert, Koen Vanhoof	2016	Fraud is a million-dollar business and it is increasing every year. Both internal and external fraud present a substantial cost to our economy worldwide. A review of the academic literature learns that the academic community only addresses external fraud and how to detect this type of fraud. Little or no effort to our knowledge has been put in investigating how to prevent and to detect internal fraud, which we call 'internal fraud risk reduction'. Taking together the urge for research in internal fraud and the lack of it in academic literature, research to reduce internal fraud risk is pivotal. Only after having a framework in which to implement empirical research, this topic can further be investigated. In this paper we present the IFR² framework, deduced from both the academic literature and from current business practices, where the core of this framework suggests to use a data mining approach.
2	Comprehensive Survey of Data Mining-based Fraud Detection Research	Clifton Phua1, Vincent Lee1 , Kate Smith1 & Ross Gayler	2010	This survey paper categorizes, compares, and summarizes from almost all published technical and review articles in automated fraud detection within the last 10 years. It defines the professional fraudster, formalizes the main types and subtypes of known fraud, and presents the nature of data evidence collected within affected industries. Within the business context of mining the data to achieve higher cost savings, this research presents methods and techniques together with their problems. Compared to all related reviews on fraud detection, this survey covers much more technical articles and is the only one, to the best of our knowledge, which proposes alternative data and solutions from related domains.
3	A fraud detection approach with data mining in health insurance	Melih Kirlidoga,b*, Cuneyt Asukb	2007	Fraud can be seen in all insurance types including health insurance. Fraud in health insurance is done by intentional deception or misrepresentation for gaining some shabby benefit in the form of health expenditures.

				Data mining tools and techniques can be used to detect fraud in large sets of insurance claim data. Based on a few cases that are known or suspected to be fraudulent, the anomaly detection technique calculates the likelihood or probability of each record to be fraudulent by analyzing the past insurance claims. The analysts can then have a closer investigation for the cases that have been marked by data mining software.
4	Computational IntelligenceModels for Insurance FraudDetection: A Review of aDecade of Research	Amira Kamil Ibrahim Hassan1 &Ajith Abraham	2018	This paper presents a review of the literature on the application of data mining techniques for the detection of insurance fraud. Academic literature was analyzed and classified into three types of insurance fraud (automobile insurance, crop insurance and healthcare insurance) and six classes of data mining techniques (classification, regression, clustering, prediction, outlier detection, and visualization). The findings of this review clearly show that automobile insurance fraud detection have also attracted a great deal of attention in recent years. The main data mining techniques used for insurance fraud detection are logistic models, Decision tree, the Naïve Bayes, and support vector machine.
5	Data Mining Techniques in Fraud Detection	Rekha Bhowmik	2006	The paper presents application of data mining techniques to fraud analysis. We present some classification and prediction data mining techniques which we consider important to handle fraud detection. There exist a number of data mining algorithms and we present statistics-based algorithm, decision tree- based algorithm and rule-based algorithm. We present Bayesian classification model to detect fraud in automobile insurance. Naïve Bayesian visualization is selected toanalyze and interpret the classifier predictions.
6	Balancing the Insurance Equation:Understanding the Climate forManaging Consumer Insurance Fraud and Abuse.	Lesch and Baker	2013	This paper by Lesch and Baker (2013) explores the landscape of purchaser insurance fraud and abuse management. It examines the elements that influence insurance fraud, consisting of social attitudes, regulatory frameworks and industry practices. By inspecting the complexities of insurance fraud detection

				and prevention, the authors spotlight the need for a balanced technique that considers each prevention and customer support. This looks at emphasizes the importance of know-how the multifaceted dynamics of coverage fraud to successfully mitigate dangers and shield the interests of insurance corporations and policyholders.

III. METHODOLOGY

The main goal of this project is to find a model that can be used to check whether certain insurance claims are genuine. This model was created after testing several algorithms to find the best model that can detect whether the claim is false or not. This is aimed at helping insurance companies create models that are more suitable for their own systems. The model should be simple enough to calculate large databases, but complex enough to obtain a high percentage of success.

Our proposed system comprises three main components: data collection, analysis, and decision-making. Firstly, data from diverse sources such as medical records, claims databases, and external sources are collected and stored in a centralized repository. Subsequently, advanced BDA techniques are employed to process and analyze this vast and heterogeneous dataset. This includes data preprocessing, feature selection, and transformation to extract valuable insights. Finally, ML algorithms are deployed for pattern recognition and anomaly detection, enabling the identification of potentially fraudulent activities

Data pre-processing

Data pre-processing in machine learning can be an important step that can make a difference in improving the quality of information to facilitate the extraction of meaningful knowledge from the information. Data preprocessing in machine learning refers to the method of preparing (cleaning and organizing) raw data to make it suitable for building and training machine learning models. After collecting the raw data, it is time to organize it so that it can be used for further processing.

Data Cleaning:

It is important that the data set is free of defects that could prevent testing or, more seriously, lead to insufficient analysis. These deficiencies or problems caused by redundant records, missing values, or loss of dimension must be effectively resolved. So, in this step bad data will be removed, and missing data will be added. The information we currently have is a comprehensive general information from which we need to remove unnecessary information and perhaps add the missing information.

Data Transformation:

In this phase data will be organized or managed so that it will be helpful to achieve the required goal.

Data Exploration:

Using Pandas data frames, the required data is extracted from the initial loaded data. Data is completely examined and discovers information and at the end concludes it to generate the report. The data is displayed using line graphs for the user to analyze using matplotlib and seaborn. For simpler comparison, all of the line graphs are displayed in one graph at the end. Prior to being combined into one large line graph for analysis and comparison with the same date, the data is first displayed on distinct graphs to demonstrate trends in the various dataset values obtained using data visualization

DOI: 10.48175/IJARSCT-23725

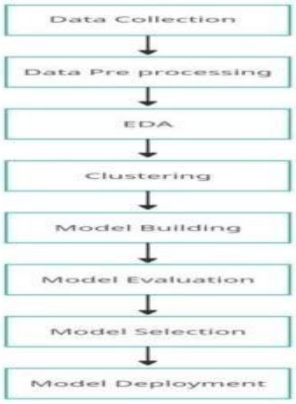


FIGURE 1. Data analysis steps

Build & training model:

Once a pattern has been identified, a suitable model needs to be constructed. A model is created by studying, practicing, and then using it. The model will be put to use and produce fraud detection. We went through these 6 processes to create and train the model:

- Contextualize machine learning in your organization
- Explore the data and choose the type of algorithm
- Prepare and clean the dataset
- Split the prepared dataset and perform cross validation
- Perform machine learning optimization
- Deploy the model

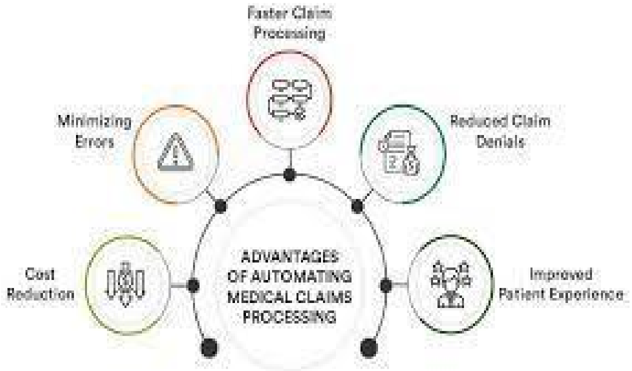

Faster Claim Processing
Minimizing Errors
Reduced Claim Denials
Cost Reduction
ADVANTAGES OF AUTOMATING MEDICAL CLAIMS PROCESSING
Improved Patient Experience

FIGURE 2. The advantages of implementing automated medical claims processing.

PROCEDURE

The platform is implemented using HTML CSS for user interface(Gui) i.e. frontend and Flask for backend and .ipynb(interactive python notebook) for the ml model designing and training with the help of python modules stated below:

sklearn:

In python, sklearn is a machine learning package which include a lot of ML algorithms.

Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy score.

NumPy:

It is a numeric python module which provides fast math functions for calculations.

It is used to read data in NumPy arrays and for manipulation purpose.

Pandas:
Used to read and write different files.
Data manipulation can be done easily with data frames.

Matplotlib:
Data visualization is a useful way to help with identify the patterns from given dataset.
Data manipulation can be done easily with data frames
There are few other modules which are used in in the training of the model to generate a better results.
Here is a list of few supervised and unsupervised techniques which are used for the system model architecture design:

Random Forest:
Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It operates by bagging (bootstrap aggregating) and random feature selection to create diverse trees, which collectively produce more robust and accurate predictions. Random Forest is widely used due to its robustness against overfitting and ability to handle high- dimensional data with ease.

Naïve Bayes Classifier:
Naive Bayes is a simple but powerful probabilistic classifier with a strong assumption of independence between features based on Bayes' theorem. Despite its weaknesses, it is generally successful in practice for information distribution and spam filtering tasks. Naive Bayes assigns the most likely class label to the input data by calculating the probability of each class given a set of input features. Its efficiency and scalability make it suitable for large, high-dimensional databases.

Support Vector Classifier(svc):
Support vector classification, or SVC, is a supervised learning algorithm used for classification problems. It works by efficiently searching a hyperplane that classifies the classes in the feature space. The goal of SVC is to maximize the distance between support vectors (data points closest to the decision boundary) of classes. With an appropriate kernel function, SVC can handle nonlinear decision boundaries and is versatile for classification problems.

Gradient boosting:
Gradient progression is a cluster learning technique that generates robust predictive models by sequentially combining weak learners (usually decision trees) with each successive model, correcting and generating errors in previous models by reducing the loss function through gradient-descent optimization. Gradient growth is known for its high prediction accuracy and is widely used in a variety of areas including regression, classification, and rank functions but can be computationally expensive and vulnerable to poor design it will be too appropriate.

Training and Testing Phase:
In this analysis, several factors were known which canfacilitate to spot for associate degree correct distinction between fraud transactions and non-fraudulent transactions that helps to predict the presence of fraud within the given transactions. Once completely different input datasets are used the Machine Learning models performed at variable performance levels. By considering average F1 score, model rankings are obtained. Higher the F1 score, higher the performance of the model. The analysis indicates that the Adjusted Random Forest formula and changed random below sampling formula provides best performance models.

IV. RESULTS

In the system, accuracy alone may not be an appropriate metric to evaluate the performance of the system. This is because accuracy measures the proportion of correct predictions made by the system out of all predictions, which may not reflect the true effectiveness of the system. Therefore, instead of accuracy, precision, recall, and F1 score are commonly used evaluation metrics for proposed systems.

For training and testing, we split the dataset 80-20 using the sklearn library. For training we fought or trained the model with 98% accuracy for testing we got 60% accuracy.

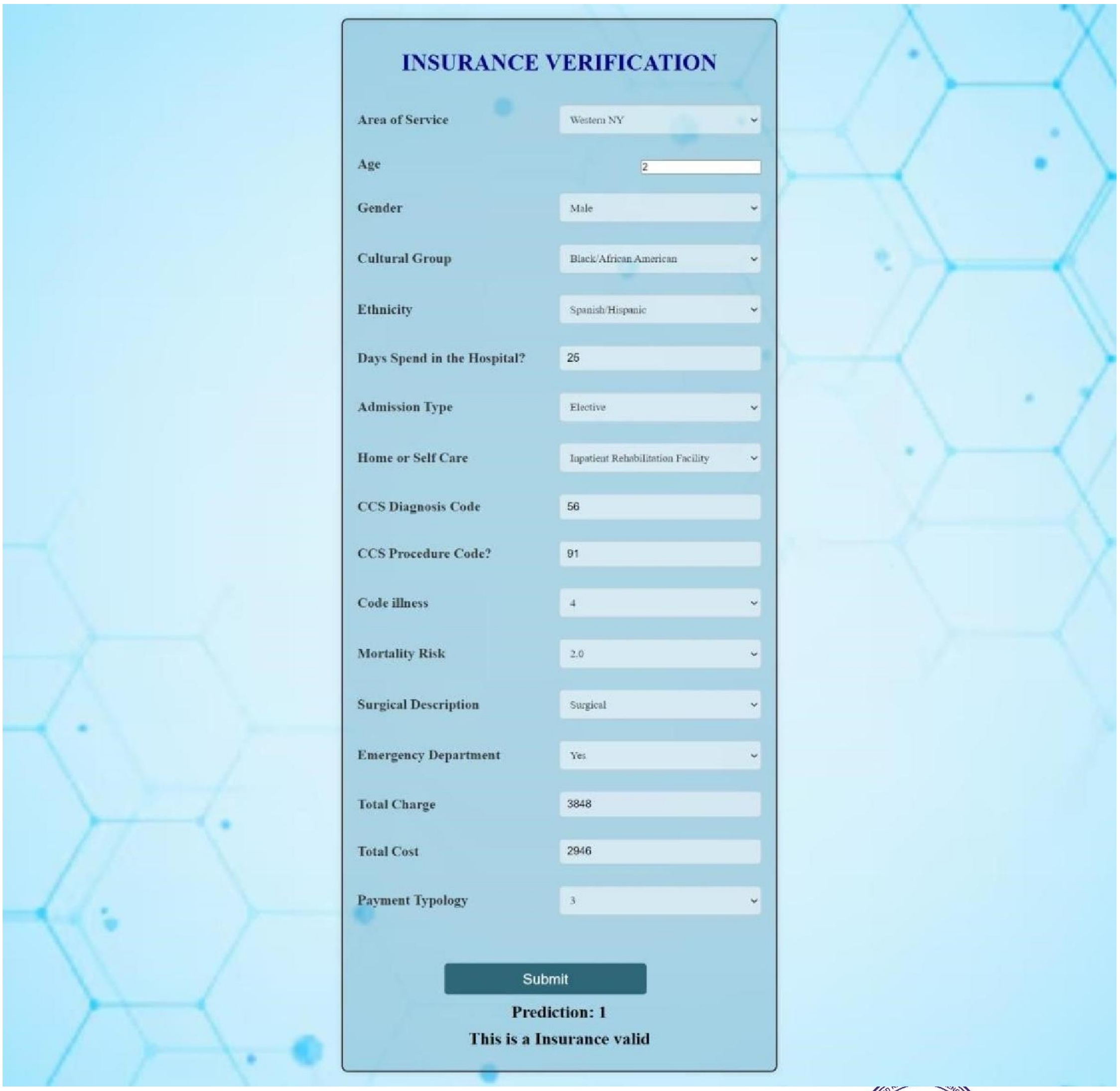
By comparing the following factors result we selected the best suitable machine learning algorithm for the training of our proposed system's model

Accuracy is defined as the percentage of correct prediction for the test data. The overall accuracy is calculated using a confusion matrix. The precision, recall and f1-score is used to calculate the overall precision. We obtained 60% accuracy of our model Overall accuracy calculated by random forest classifier.

Precision: The precision measures the fraction of recommended problems that were actually relevant to the user. It is calculated as the number of relevant recommended problems divided by the total number of recommended problems.

Recall: The recall measures the fraction of relevant problems that were recommended to the user. It is calculated as the number of relevant recommended problems divided by the total number of relevant problems.

F1 score: The F1 score is the harmonic mean of precision and recall, which provides a balanced measure of the model's overall performance. It is calculated as follows: *F1 Score = 2 * (Precision * Recall) / (Precision + Recall)*



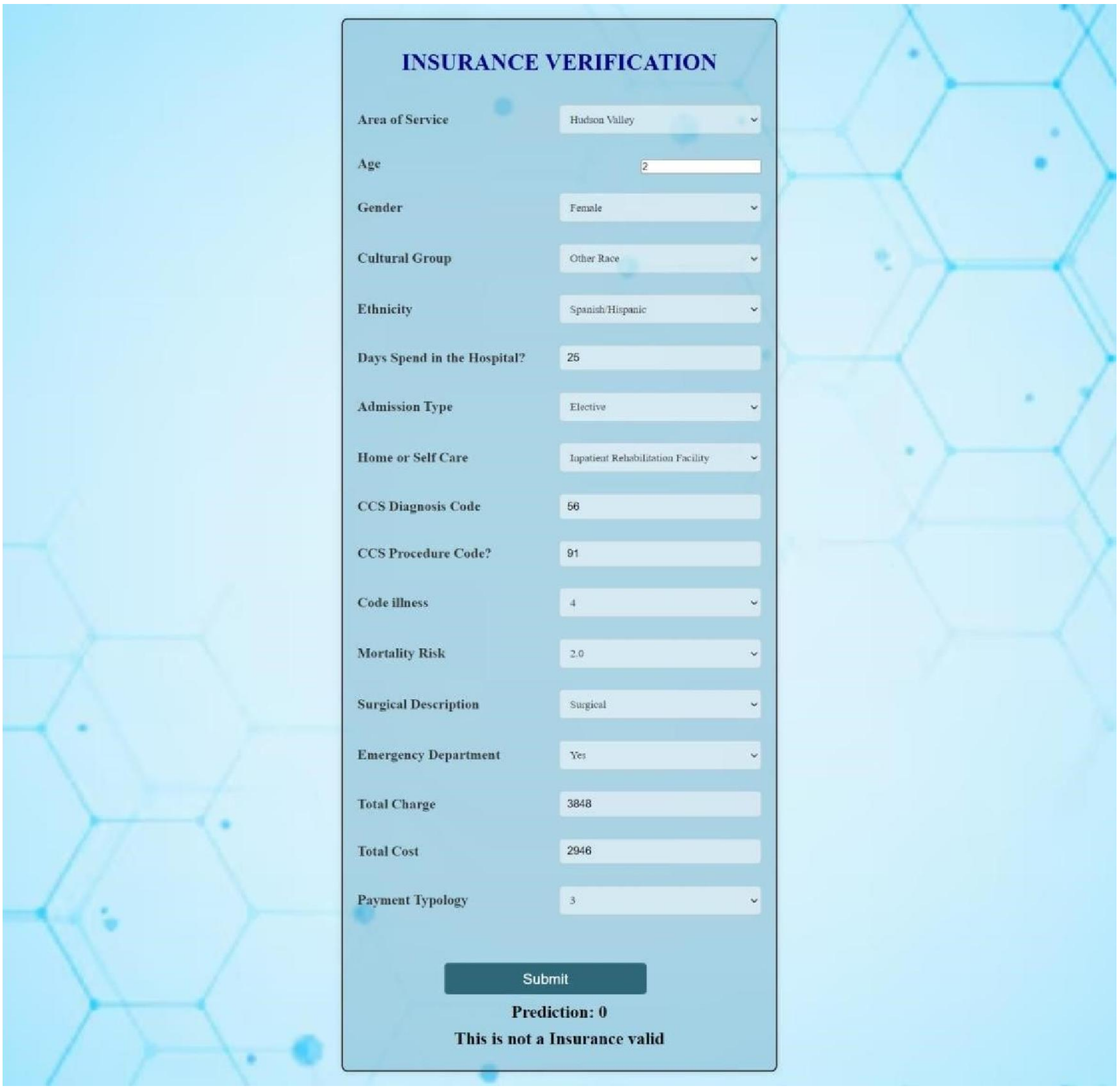
INSURANCE VERIFICATION

Area of Service	Western NY
Age	2
Gender	Male
Cultural Group	Black/African American
Ethnicity	Spanish/Hispanic
Days Spend in the Hospital?	25
Admission Type	Elective
Home or Self Care	Inpatient Rehabilitation Facility
CCS Diagnosis Code	56
CCS Procedure Code?	91
Code illness	4
Mortality Risk	2.0
Surgical Description	Surgical
Emergency Department	Yes
Total Charge	3848
Total Cost	2946
Payment Typology	3

Prediction: 1

This is a Insurance valid

DOI: 10.48175/IJARSCT-23725



INSURANCE VERIFICATION

Area of Service	Hudson Valley
Age	2
Gender	Female
Cultural Group	Other Race
Ethnicity	Spanish/Hispanic
Days Spend in the Hospital?	25
Admission Type	Elective
Home or Self Care	Inpatient Rehabilitation Facility
CCS Diagnosis Code	56
CCS Procedure Code?	91
Code illness	4
Mortality Risk	2.0
Surgical Description	Surgical
Emergency Department	Yes
Total Charge	3848
Total Cost	2946
Payment Typology	3

Prediction: 0

This is not a Insurance valid

V. CONCLUSION

As countries around the world develop into more economic-oriented countries, the goal is to stimulate their economies. Fighting these fraudsters and money launderers was a very complex task before the era of machine learning but thanks to machine learning and AI we are able to fight these types of attacks. The proposed solution can be used to check whether a particular insurance claim made in insurance companies is valid or not. The model should be simple enough to calculate large datasets, yet complex enough to have a decent success percentage. Under this project, we select more than 1000 data samples and split the data into training and test data.

We can see that, compared to algorithms XGBoost and Random Forest algorithms; has better performance than KNN. We have just introduced the feature of Machine Learning Algorithms. We have worked with more algorithms and finally calculate which gives more precision, accuracy and recall. The integration of Machine Learning (ML) techniques with web development technologies like HTML, CSS, and Flask presents a promising solution for addressing the pervasive issue of medical insurance fraud. By leveraging ML algorithms for pattern recognition and anomaly detection, coupled with the userfriendly interface provided by HTML and CSS, our approach ensures accessibility and ease of use for both healthcare professionals and insurance providers. The incorporation of Flask, a lightweight web framework, facilitates seamless communication between the ML system and the user interface,

DOI: 10.48175/IJARSCT-23725

enabling real-time feedback and response. Through this interdisciplinary approach, we have demonstrated the potential to enhance the efficiency and accuracy of fraud detection processes in the healthcare industry. Moreover, by humanizing the interface and streamlining the user experience, our system promotes greater transparency and trust in insurance transactions. Moving forward, continued research and development in this field hold the promise of further advancements in medical insurance fraud detection, ultimately safeguarding the financial resources and integrity of healthcare systems worldwide.

VI. FUTURE WORK

The future scope for a medical insurance validation system is promising, with advancements in artificial intelligence and data analytics. Enhanced algorithms can streamline claims processing, reduce fraud, and improve accuracy in determining coverage eligibility. Integration with wearable devices and electronic health records could enable real-time monitoring of health metrics, leading to personalized insurance plans. Moreover, blockchain technology may offer secure, transparent transaction records. Overall, such innovations hold the potential to revolutionize the efficiency and effectiveness of medical insurance validation systems.

REFERENCES

[1]. S. Subudhi and S. Panigrahi, “Detection of Automobile Insurance Fraud Using Feature Selection and Data Mining Techniques,” International Journal of Rough Sets and Data Analysis, vol. 5, no. 3, pp. 1–20, Jul. 2018.

[2]. M. Nur Prasasti, A. Dhini, and E. Laoh, “Automobile Insurance Fraud Detection using Supervised Classifiers,” 2020 International Workshop on Big Data and Information Security (IWBIS), Oct. 2020.

[3]. X. Liu, J.-B. Yang, D.-L. Xu, K. Derrick, C. Stubbs, and M. Stockdale,“Automobile Insurance Fraud Detection using the Evidential Reasoning Approach and Data-Driven Inferential Modelling,” 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Jul. 2020.

[4]. C. Yan and Y. Li, “The Identification Algorithm and Model Construction of Automobile Insurance Fraud Based on Data Mining,” 2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC), Sep. 2015.

[5]. Bodaghi and B. Teimourpour, “Automobile Insurance Fraud Detection Using Social Network Analysis,” Lecture Notes in Social Networks, pp. 11–16, 2018.

[6]. V. Jain, “Perspective analysis of telecommunication fraud detection using data stream analytics and neural network classification-based data mining,” International Journal of Information Technology, vol. 9, no. 3, pp. 303–310, Aug. 2017.

[7]. Data Mining Techniques in Fraud Detection Author: Rekha Bhowmik

[8]. A Comprehensive Survey of Data Mining-based Fraud Detection Research

[9]. A fraud detection approach with data mining in health insurance Author: Melih Kirlidoga,b*, Cuneyt Asukb

DOI: 10.48175/IJARSCT-23725