IJNRD.ORG ISSN : 2456-4184
International Research Journal IJNRD Research Through Innovation
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

FRAUDULENT HEALTH INSURANCE CLAIMS DETECTION USING MACHINE LEARNING

Ch. Lalitha Syama Sundari[1], G. Harika[2], G. Mahi Durga Lakshmi[3], J. Pavaneeth[4], G.Bhargav Ram[5]
12345 Department of Information Technology,
Sasi Institute of Technology & Engineering

Abstract : Health insurance fraud harm the integrity and long-term viability of global healthcare systems. With the expanding use of digital technology and electronic health records, there is a greater need for effective fraud detection tools to defend against financial losses and assure the quality of service. This review paper provides a comprehensive summary of current research on detecting false health insurance claims utilising machine learning approaches. This paper discusses a variety of methodology, including supervised and unsupervised learning algorithms, feature engineering techniques, anomaly detection methods, and ensemble learning approaches. It examines the issues of imbalanced datasets, noisy data, and model interpretability, as well as techniques for overcoming them.This research also assesses the effectiveness of machine learning models in detecting false health insurance claims utilizing real-world datasets and performance measurements such as accuracy, precision, recall, and F1 score. We hope that this poll will provide useful insights into the present status of research in this subject, as well as indicate future research directions to improve healthcare fraud detection systems.

Index Terms *- **Health insurance, Insurance claims Fraud detection, and Machine learning.***

I.INTRODUCTION

Health insurance acts as a safety net which pays for medical and surgical expenses incurred by the insured, providing a financial safety against the expensive costs associated with healthcare services. It guarantees access to essential medical care, including as preventative treatments, immunizations, and screenings, boosting early detection and treatment of health problems. Health insurance provides people and families with financial security and peace of mind, lessening the burden of medical bills and allowing them to concentrate on their health without fear of financial ruin. Furthermore, in some countries, health insurance is mandatory, stressing its significance in assuring access to healthcare and supporting public health programs. Health insurance policies include a wide range of coverage options designed to satisfy the different healthcare needs of people and families. These insurance often cover a variety of services, such as doctor visits, hospital stays, prescribed drugs, diagnostic testing, and preventive care. Health insurance policies are categorized into four types: Health Maintenance Organization (HMO) plans, Preferred Provider Organization (PPO) plans, Exclusive Provider Organization (EPO) plans, and Point of Service (POS) plans, with each providing varying levels of flexibility and access to healthcare providers. Policies also differ in terms of cost-sharing structures, such as deductibles, copayments, and coinsurance, as well as coverage restrictions and exclusions. Depending on the insurance plan and provider, health insurance policies sometimes include extra benefits such as dental care, visionary care, mental health services, and prescription medication coverage. Ultimately, the choice of health insurance policy is determined by individual preferences, healthcare needs, and budgetary factors, emphasizing the necessity of making informed decisions when picking a health insurance plan. Insurance claims can be filed when an insured person or policyholder suffers a covered loss or incurs qualified expenses as specified in their insurance policy. The process of claiming insurance normally begins with alerting the insurance company of the loss or expenses incurred within the timeframe set in the policy. Insurers frequently ask policyholders to submit a claim form detailing the incident or expenses, along with any supporting paperwork such as medical bills, receipts, or repair estimates. When a claim is filed, the insurance company will conduct an investigation to determine the veracity of the claim and the coverage available under the policy. Once approved, the insurer will refund the policyholder for the covered losses or expenses, either directly to the service provider or to the policyholder. To ensure a smooth and prompt claims settlement, policyholders should get familiar with the claims procedure and follow any deadlines or conditions mentioned in their insurance policy. Machine learning (ML) is a branch of artificial intelligence where algorithms and statistical models are created that allow computer machines to learn and improve accuracy through experience without explicit programming. ML algorithms are capable of analysing massive datasets, identifying patterns, and making data driven predictions or judgments. Machine learning is critical in fraud detection because fraudulent operations are complicated and dynamic. Traditional rule-based fraud detection systems generally fail to keep up with changing fraud strategies, necessitating frequent updates and manual intervention. Machine learning, on the contrary, provides a more adaptive and automated approach to fraudulent detection by constantly learning from new data and modifying its models as needed.ML algorithms can evaluate large volumes of data to find subtle patterns and abnormalities which indicate fraudulent behaviour, that allows more accurate and quick identification of fraudulent operations while decreasing false positives. This makes machine learning an effective tool for improving fraud detection capabilities in a variety of industries, including finance, insurance, healthcare, and ecommerce. In health insurance claims, there are various potential channels for fraudulent activity, which may include, but are not limited to: charging for services not rendered, which is known as phantom billing, where healthcare providers submit claims for procedures or treatments that never took place; upcoding, which involves inflating the cost of services and medical procedures to receive higher reimbursements; unbundling, where services that should be billed together are separately billed to maximize reimbursement;

and falsifying patient information or medical records to support. Besides, cooperation between healthcare providers and patients, such as kickback schemes or superfluous referrals and treatments, can lead to fraudulent activity in health insurance claims. Identifying and mitigating these various types of fraud is critical to maintaining the integrity of the health insurance system and guaranteeing fair and accurate reimbursement for legitimate healthcare services. Detecting fraudulent health insurance claims is crucial for insurance companies and healthcare service providers to maintain financial integrity and ensure equitable resource distribution. Traditional manual approaches for detecting fraudulent activity are frequently time-consuming, labour intensive, and subject to human error. However, as machine learning algorithms have progressed, there has been a paradigm shift toward automated fraud detection systems capable of efficiently and accurately analysing large volumes of healthcare data. Machine learning algorithms can adapt to patterns, anomalies, and suspect behaviours in health insurance claims data, allowing for the rapid identification and mitigation of fraudulent activity. Machine learning enhances fraud by utilizing supervised and unsupervised learning methodologies, predictive modelling, feature engineering, and ensemble methods. Increasing detection capabilities, lowering false positives, and adjusting to changing fraud methods. This study examines the utilization of machine learning to detect fraudulent health insurance claims, emphasizing its importance in enhancing fraud detection accuracy, efficiency, and overall operational performance in the healthcare insurance business

II. Related Work

After studying relevant literature, we learn about research efforts and data mining approaches used to detect insurance fraud in healthcare data. Jing Li [6] conducted a comprehensive survey of statistical methods used to detect healthcare fraud. The study focused on classifying fraudulent behavior, identifying major sources of fraud detection, and comparing statistical fraud detection methods to determine future research direction;Joudaki [10] conducted a comprehensive study on data mining and machine learning strategies for detecting healthcare fraud. Joudaki discusses supervised and unsupervised learning techniques and recommends broad procedures for mining healthcare claims data. Travaille [12] studied the effectiveness of existing electronic fraud detection tools in businesses similar to the US Medicaid program. While there is limited research on detecting insurance fraud, it is clear that devoted efforts are required for health insurance claims. This section highlights ongoing research in the health service realm. Ortega [7] suggested a medical claim fraud detection system for a Chilean private health insurance business, using multi-layer perceptron neural networks for each entity implicated in fraud. The results suggest a monthly detection rate of 75 frauds. Yang et al, [8] present a data mining methodology that uses clinical pathways to build scalable and extensible detection models. The proposed methodologies were evaluated using real-world data from Taiwan's National Health Insurance (NHI) program. Empirical trials demonstrate that the detection approach is effective in identifying fraudulent and abusive cases that may otherwise go undetected. In 2007, Kuo et al, introduced a new data mining method for the National Health Insurance Bureau of Taiwan Government's medical database [9]. The database is clustered using the ant system-based clustering technique and ant k-means, followed by the ant-colony system-based association rules mining algorithm to identify valuable rules for each group. The results showed that the proposed approach extracts rules more quickly and identifies essential ones. Thornton [11] advocated using unsupervised learning and outlier detection to detect fraud in Medicaid claims. The metrics created for dental practitioners were tested using outlier detection, resulting in successful identification of fraudulent activity. Qi Liu's [13] suggested health care fraud detection system uses a clustering algorithm with geo-location information from both Medicaid beneficiaries and providers to identify questionable claims. Kose et al [14] propose detecting fraudulent activities independent of the people and commodities involved in claims. To detect fraud, an interactive machine learning methodology is used. Methods include pair-wise comparison, analytic hierarchical processing, and expectation maximization to cluster comparable actors. Validation is built on realworld data, and area under curve figures are given for each experiment. Thorton et al [15] suggested multidimensional data models and analysis tools to predict fraudulent activity. The approach is examined for Medicaid and US healthcare. Fraud is a far-spread issue that affects all domains. Fraud detection systems are always evolving, and thieves are developing new strategies to counter them. Machine learning proved to be more effective than manual auditing for detecting fraud. The literature evaluation suggests that both supervised and unsupervised learning techniques can be utilized to detect fraud. Various methods, such as supervised and unsupervised learnings, have been used to detect fraud. Various techniques, including supervised methods such as naive Bayes, neural networks, and logistic regression, and unsupervised techniques like clustering and outlier identification, have been utilized to detect fraud in the literature. Among these neural networks was demonstrated that this approach is highly effective for non-linear data structures. Richard Bauder and his team [16] conducted extensive research on fraud detection using supervised and also unsupervised learning methods. The researchers began by analyzing data from part B providers and utilizing outlier detection techniques. They combined Medicare-related datasets and labeled them using the LEIE exclusion list. They addressed data imbalance issues using diverse approaches at the data and algorithmic levels. According to their analysis, data-level approaches outperformed algorithmic approaches. Panigrahi combined Dempster-Shafer and Bayesian learning algorithms to filter and generate scores. Next, they utilized a Bayesian learner to detect fraud if Dempster-Shafer did not recognize it. Some efforts have attempted to integrate AI and data mining approaches. The artificial immune system was used to detect fraud in the their studies. These approaches are adaptable to evolving fraud tendencies. [10] In his paper, Ibtissam Benchaji used clustering techniques like kmeans and a genetic algorithm to address imbalanced class distributions during credit card fraud detection. Alamelu et al (2011) analyzed the financial success of Indian insurance companies, focusing on asset requirements and management systems [16]. Tkachenko et al (2018) suggested a piecewise-linear strategy for predicting medical insurance costs, utilizing the SGTM neural-like structure. They trained multi-layer perceptron and common SGTM neural-like structures. The researchers gathered data on insurance costs in various locations of the United States. The proposed approach had a MAPE percentage of 30.60400373 and MAE of 3453.293634, indicating excellent performance [17]. Shinde et al. (2020) investigated several regression and neural network models along with Support Vector Machine, Multiple Linear Regression, XGBoost, Random Forest Regressor, and Deep Neural Networks. It has been discovered that Deep Neural Networks were the most effective strategy. With an RMSE of 0.0695 and an accuracy of 87.95 [18]. Chowdhury et al (2020) combined Neural Network (NN) and Internet of Things (IoT) to estimate health insurance costs. The proposed Fitness-Dependent Randomized Whale Optimisation Algorithm (FR-WOA) was the most effective [19]. Izonin et al (2019) developed a new method for building a committee using SGTM neural architecture and an RBF kernel. They utilized the SGTM neural architecture, multilayer perceptron, adaptive boosting, and stochastic gradient descent regressor. Bayerstadler et al. [14] developed a predictive algorithm for detecting fraud and abuse that used manually labeled claims as training data. The approach predicted fraud and abuse scores using a probability distribution for new claim invoices. The authors suggested a Bayesian network to summarize medical claim representation patterns through latent variables. They also used MCMC to estimate model parameters [15]. Zhang et al (16) presented a Medicare fraud.

Yang and Hwang created a healthcare fraud detection algorithm based on clinical pathways and a process mining architecture [13]. The approach utilizes a module to identify structural patterns in both good and negative clinical occurrences. The module extracted frequent patterns from every clinical case. Next, a feature-selection module filters the dataset and labels its features. An inductive model is developed using the feature set to evaluate new claims. Their method involves clustering, association analysis, and principal component analysis. The approach was applied to real-world data from Taiwan's National Health Insurance (NHI) program. The writers used several elements to create patterns for both normal and the importance of these characteristics is not examined in relation to abusive accusations.

Table 1: Techniques used by different Authors

Author name	SVM	DT	RF	NB	CNN	LR	GB	KNN	K-C	DL	ARM	MLP	PR	ANN	PCA
Kritika Dutta		✓	✓			✓							✓		
P Naga Jyothi[1]															
Jin-Su Park															
Shamitha S.K	✓			✓		✓			✓			✓			✓
Dr Aliza Binit Sarlan	✓										✓				
Sharifa Rigga Mambo		✓		✓				✓							
Abdullah Alanazi[2]			✓			✓								✓	
Lavanya Settipalli[3]															
Venkateswarlu Nalluri	✓	✓	✓									✓			
Aayushi Verma											✓				
Thomas George K		✓	✓												
Richard A. Bauder			✓	✓			✓	✓							
Shashank Agarwal[1]	✓	✓				✓									
Paresh Gohil	✓	✓	✓	✓				✓							
Laiqa Rukhsar[1a]	✓	✓	✓	✓		✓		✓				✓			
Anokye Acheampong Amponsah		✓													
Pallav Kumar Baruah[3]					✓										
Robert A. Sowah ,[1]	✓														
Arif Ismail Alrais		✓													

Research Questions :
[1] :What are the most effective machine learning strategies and methodologies for detecting fraudulent activities within health insura nce claims data?
[2] :How can machine learning models address challenges such as data imbalance, where fraudulent instances are significantly outnumber ed by legitimate claims?
[3] :What approaches can be employed to adapt machine learning models to evolving fraud schemes, ensuring co ntinuous effectiveness in fraud detection?

III. METHODOLOGY

1.Data Preprocessing: Data pretreatment for detecting fraudulent health insurance claims entails cleaning the dataset by removing duplicates, addressing missing values, and fixing errors. To prevent dominance during model training, numerical features are normalized or standardized, whereas categorical variables are encoded as numerical representations using approaches such as one-hot encoding or label encoding. Class imbalance is resolved by oversampling, underSampling, or algorithms intended to handle unbalanced data, such as SMOTE.
2.Feature Extraction: Feature extraction is the process of obtaining useful information from a set of data that can be utilized to detect fraudulent health insurance claims. This contains both claim-related characteristics, such as claim amount, kind of procedure, and length of hospital stay, and patient-related features, such as age, gender, pre-existing conditions, and medical history. Provider information, such as specialty and location, as well as temporal variables like day of the week or season, are extracted to supplement the dataset with useful information.
3.Machine Learning Algorithm: Because labelled data is readily available, supervised machine learning algorithms are frequently used to detect fraudulent health insurance claims. Random Forest Algorithm, Decision Tree, Gradient Boosting Machines (GBM), and Support Vector Machine Algorithm (SVM) are popular alternatives due to their capacity to handle complex relationships and large amounts of data. Unsupervised learning algorithms, such as clustering and anomaly detection, can also be investigated, especially in cases where labelled data is scarce or unavailable.
4.Prediction: After training the machine learning model on pre-processed data, it is applied to new, previously unknown insurance claims data to determine if each claim is fraudulent or real. Predictions are made using the model's output, which may include probabilities or class labels indicating the possibility of fraud. This stage is critical for detecting suspected fraud and applying appropriate risk mitigation measures. 5.Performance Comparison: Performance comparison includes assessing the trained machine learning algorithms using metrics such as precision, recall, F1-score, accuracy, ROC-AUC, and confusion matrix. The performance of many models is evaluated to determine which one of the algorithms best fulfills the insurance company's needs, taking into considerations such as false positive and false negative rates. Cross-validation will be used to test the models' robustness and ensure consistent performance across diverse subsets of the data.

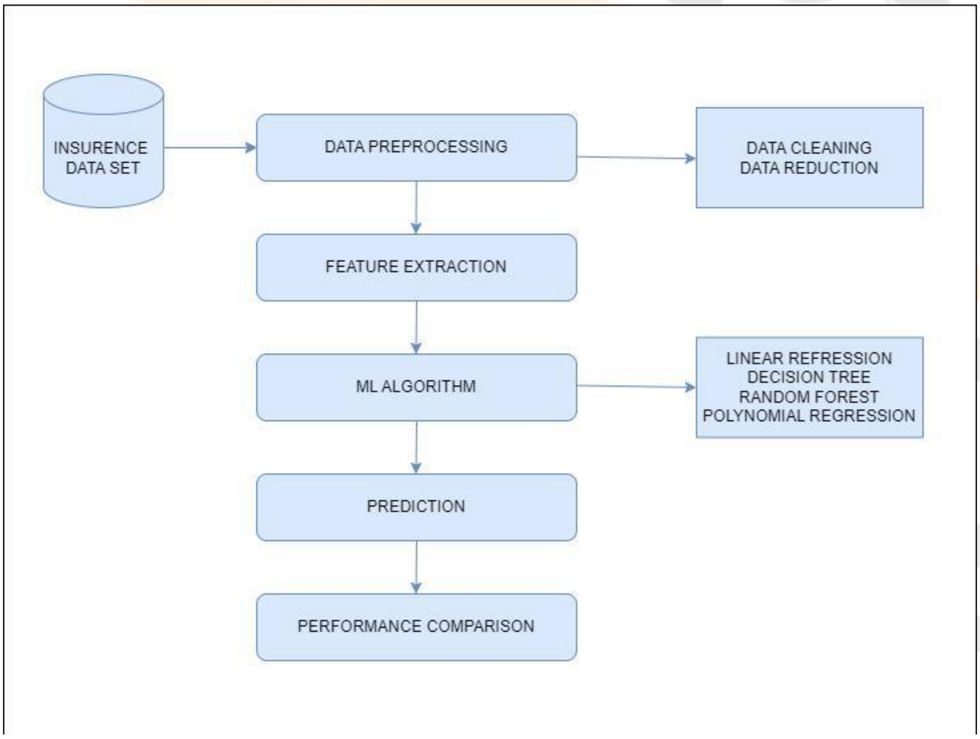

Fig 1: Methodology

IV.ALGORITHMS

1. **Linear regression**: Linear regression is among the popularly used supervised learning approaches that predicts continuous outcomes. It approximates the relationship between one dependent variable and one or multiple independent variables by fitting a linear equation to the observed data.

Formula: $y= a_0+a_1x+ \varepsilon$

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

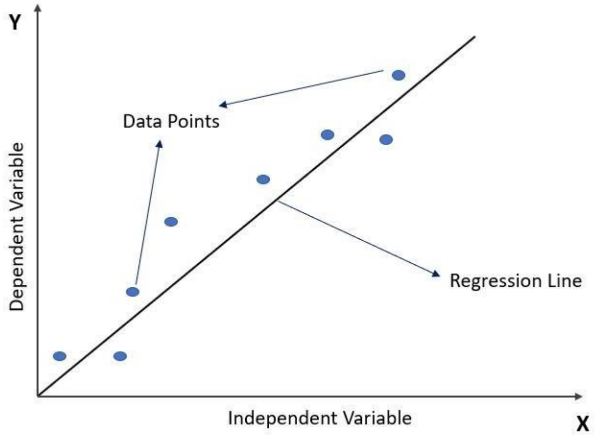

Fig 2: Linear Regression

Linear regression models the relationship between one dependent variable and one or more than one independent variable. It presupposes a linear relationship between the input features and the goal variable. The model calculates the linear equation's coefficients to minimize the sum of the squared discrepancies between observed and forecasted values.Linear regression is commonly utilized for prediction and inference applications because it is simple and easy to interpret.However, it may miss intricate relationships in the data and is vulnerable to outliers.Regularization techniques like as Lasso and Ridge regression can help to reduce overfitting.It is appropriate for situations where the relationship between variables is roughly linear.Linear regression is less good at detecting nonlinear correlations.Model performance is typically evaluated using measures such as Mean Squared Error (MSE) and R- squared. It is utilized as a basis model for more advanced algorithms such as logistic regression and decision trees.

2. Logistic regression: Logistic regression is a prominent supervised learning technique for binary classification problems. It estimates the probability of a binary outcome by fitting a logistic function to the observed data, making it appropriate for situations involving a categorical dependent variable with two levels.

Formula: S(z)=1/1+e−z

Here, z represents the linear combination of input features and their associated weights, often denoted as θTx, where θ is the vector of weights, and x is the vector of input features. The hypothesis function hθ(x) in logistic regression is then given by applying the sigmoid function to z: hθ(x)=S(θTx) = 1/1+e−θTx

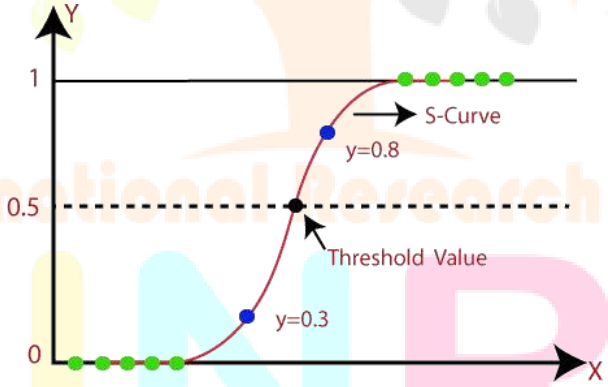

Fig 3: Logistic Regression

Logistic regression is a binary classification procedure that calculates the likelihood of a binary result using one or more predictor variables.It calculates probabilities using the logistic function, often known as the sigmoid function.Logistic regression is based upon the assumption that there is a linear relationship between the target variable's log chances and the input features.It is commonly utilized in industries such as healthcare, finance, and marketing for binary classification. Logistic regression produces probabilities that can be thresholded for binary predictions.Regularization techniques such as Lasso and Ridge regression can be used to avoid overfitting. Performance is measured using metrics such as accuracy, precision, recall, and Area Under the ROC Curve (AUC-ROC).Logistic regression is interpretable and can provide information regarding the factors impacting the outcome.It is appropriate for situations where the relationship between input features and goal variables is roughly linear.Logistic regression can handle outliers in the input data.

3. Support Vector Machines (SVMs): Support Vector Machines are sophisticated supervised learning algorithms used in classification and regression. In binary classification, SVM determines the hyperplane that best separates data points from distinct classes, maximizing the margin between the hyperplane and the support vectors. SVM can handle both linear and non-linear decision boundaries by employing various kernel functions, including linear, polynomial, and radial basis function (RBF) kernels. SVM strives to reduce classification error while increasing margin, making it resistant to overfitting and appropriate for high- dimensional data.

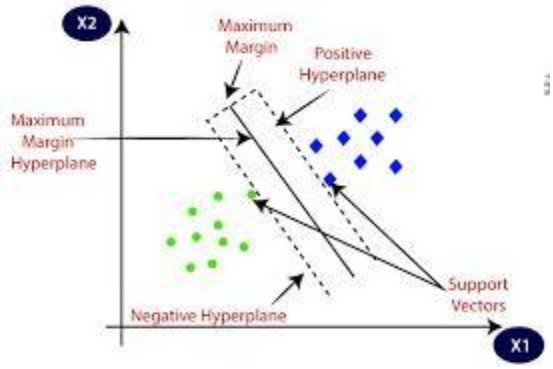

Fig 4: Support Vector Machine

The Support Vector Machine algorithm (SVM) is a strong supervised learning technique used for classifying tasks. It determines the best hyperplane that separates data points from distinct classes with the greatest margin. SVM can handle high-dimensional data and is useful in circumstances where the data cannot be separated linearly by transforming the input space with kernel functions. The kernel function (e.g., linear, polynomial, radial basis function) influences the model's performance. SVM is resistant to overfitting, particularly in high-dimensional domains, thanks to the margin maximization target. It is widely applicable across a variety of fields, including healthcare, finance, and text classification.SVMs can perform both linear and nonlinear classification tasks. Performance is measured using metrics like as accuracy, precision, recall, and Area Under the ROC Curve (AUC-ROC). SVM can be computationally expensive, particularly with large datasets, and hyperparameters may need to be carefully tuned. SVM works well for binary classification tasks, but it may need to be extended for multiclass classification.

4:Decision Trees: Decision Trees, which are versatile supervised learning algorithms used for classification techniques and regression. They recursively partition the feature space into smaller regions depending on feature values and predict using the decision rules at each node. Decision trees are interpretable and can handle numerical and categorical data, making them useful in various applications. A decision tree's splits are determined using metrics such as information gain or Gini impurity, with the goal of maximizing target variable homogeneity within each partition while minimizing impurity between partitions.

Formula: The Gini impurity is a measure of how often a randomly chosen element would be incorrectly classified. For a binary classification problem with classes C1 and C2, the Gini impurity (IG) is calculated as:

IG=1−∑i=1npi2

where pi is the probability of belonging to class Ci. The goal is to minimize the Gini impurity when making splits in the data**.**

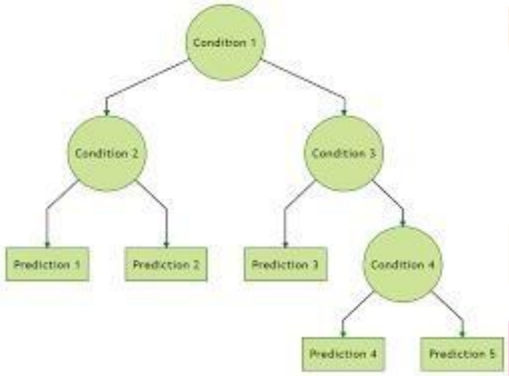

Fig 5: Decision Tree

Decision Tree is a supervised learning system that can do both classification and regression tasks. It divides the feature space into disjoint sections and predicts the target variable's value using the majority class (classification) or average (regression) within each region. The decision tree splits the data recursively depending on feature thresholds in order to optimize information gain (classification) or decrease impurity. It is interpretable and can detect complex nonlinear relationships in data. Decision trees are often prone to overfitting, especially deep trees, and may require pruning strategies to generalize successfully on previously encountered data. The algorithm is susceptible to outliers and may produce biased trees in the presence of imbalanced classes. Decision Trees are appropriate for datasets containing both numerical and category variables. Performance is assessed using.

5. Random Forests: Random Forests are ensemble learning techniques that mix numerous decision trees to increase forecast performance while minimizing overfitting. Each decision tree in the random forest is trained independently, with a bootstrapped subset of the training data and a random subset of features at each split. The final prediction is obtained by aggregating the predictions of all individual trees, either by average (regression) or voting (classification). Random forests are stable, can handle high- dimensional data well, and are less prone to overfitting than individual decision trees, making them suited for an extensive range of machine learning tasks.

Formula: RF(X)=1/N

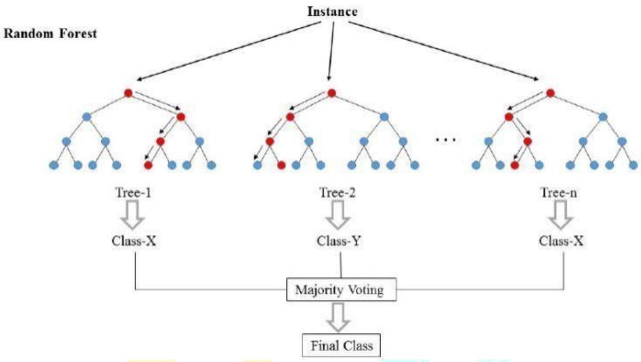

Fig 6: Random Forest

Random Forest can be defined as a learning technique that generates several decision trees during training and returns either the mode of the classes in case of classification or the mean prediction regression of each tree in case of regression. It incorporates rando mization into the tree-building process by taking a random subset of characteristics at each split and bootstrapping samples from the training dataset. Random Forest is resistant to overfitting and noise in data, making it ideal for high-dimensional datasets. It can handle classification and regression problems and is best capable at capturing complex nonlinear interactions. Random Forest calculates feature importance scores, which help users to determine the relative value of input features. The algorithm is less prone to overfitting than individual decision trees.Random Forest is widely utilized in a variety of applications, including healthcare, finance, and bioinformatics. Performance is measured using metri cs like as accuracy, precision, recall, and Area Under the ROC Curve (AUC-ROC). Random Forest is computationally efficient in both training and inference, making it ideal for huge datasets. The number of trees present in the forest and the maximum depth of the trees are critical hyperparameters that influence model performance.

6. Gradient Boosting: Gradient Boosting is an ensemble learning strategy for regression and classification tasks that creates a strong predictive model by gradually adding weak learners to the ensemble. Gradient boosting iteratively constructs the ensemble by fitting a w eak learner, usually a shallow depth decision tree, to the preceding model's residuals. The learning procedure consists of maximizing a differentiable loss function by minimizing the residuals at each iteration, which updates the weak learner's parameters. Gradient boosting met hods, such as Gradient Boosting Machines (GBM) and XGBoost, are well-known for their superior prediction accuracy and ability to detect complicated patterns in data.

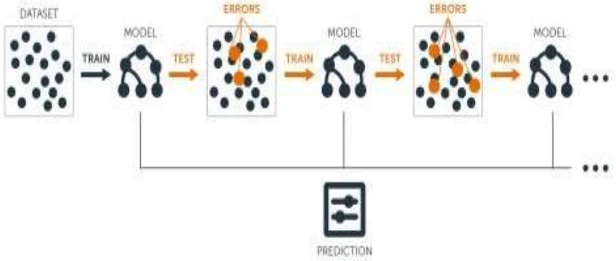

Fig 7: Gradient Boosting

Gradient Boosting is an ensemble learning strategy that systematically combines numerous weak learners, such as decision trees, to create a powerful prediction model. It reduces the model's error by optimizing a differentiable loss function with gradient descent. Gradient Boosting adds decision trees to the ensemble in a progressive manner, with each tree focused on the previous tree's faults. It is resistant to overfitting and may detect complex nonlinear correlations in the dataset. Gradient Boosting is widely used in applications such as XGBoost, LightGBM, and CatBoost. Gradient Boosting is frequently applied in a various fields, including healthcare, finance, and marketing.It can perform regression tasks and classification tasks. Performance is measured using measures such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Area Under the ROC Curve (AUC-ROC). Gradient boosting is computationally intensive and may necessitate the adjustment of hyperparameters such as learning rate, tree depth, and regularization parameters. It frequently outperforms the industry standard in predictive modeling jobs.

7.K-Nearest Neighbors (KNN):

K-Nearest Neighbors is defined as a straightforward supervised learning technique used for classification and regression tasks. KNN generates predictions by determining the majority class (classification) or averaging the values (regression) of the K nearest data points in the feature space. The distance metric used to calculate the nearest neighbors varies, with Euclidean distance being a popular choice. KNN is a non-parametric and instance-based method, which means it does not make explicit assumptions about the underlying data distribution and instead depends on training data for prediction. Formula: Let X be the feature space, and xi and xj be two instances in X. The distance between xi and xj is often calculated using a distance metric, such as Euclidean distance or Manhattan distance. The general formula for Euclidean distance is:

Distance (xi,xj)=(∑k=1n(xik−xjk)2)1/2

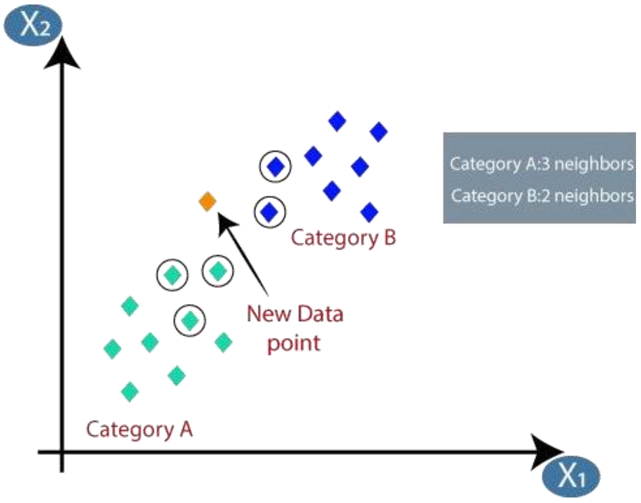

Fig 8: K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a straightforward technique utilized for classification and regression. It produces predictions using the majority class or the average of the k closest data points in the feature space. KNN is a non- parametric method, which means it makes no assumptions about the underlying data distribution. It is resistant to noisy data and can deal with complex decision boundaries. The performance of KNN is strongly subjected to the distance metric used and the value held by k. The algorithm must save all training data points, which makes it memory- intensive for large datasets.KNN is sensitive to the size of input features, and feature scaling is frequently required for peak performance. It is appropriate for datasets with small dimensions and can be used as a benchmark model for comparison. Accuracy, precision, recall, and F1-score are commonly used measures for evaluating KNN performance. KNN is computationally expensive during inference, particularly for large datasets, because it must calculate distances to all training point.

Table 2: Comparative Analysis

S.NO	AUTHORS	METHOD	ACCURACY	APPROACH
1	Richard A. Bauder	Compare several machine learning methods for detecting Medicare fraud, including supervised, unsupervised, and hybrid approaches.	83%	Utilize the 2015 Medicare Provider Utilization and Payment Data along with the List of Excluded Individuals/Entities database.
2	Paresh Gohil	Discuss the utilization of machine learning techniques for fraud detection in the medical insurance claim system, including Support Vector Machine, K-nearest Neighbor, Random Forest, Decision Tree, and Navier Bayes algorithms.	89%	Propose a framework based on machine learning algorithms and mention the potential effectiveness of automated approaches.
3	Anokye Acheampong Amponsah	Propose a novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology, utilizing a decision tree classification algorithm.	92%	Utilize machine learning strategies and blockchain to detect and prevent fraudulent healthcare claim processing.
4	Rohan Yashraj Gupta	Conduct a comparative study in evaluating the effectiveness of various machine learning and deep learning-based fraud detection models for Universal Health Coverage schemes.	86%	Apply various machine learning strategies and deep learning models to the dataset from Ayushman Bharat (PM-JAY) and evaluate their performance.
5	Saba Kareem	Propose a framework for identifying fraud health insurance claims utilizing Associate Rule Mining combined with supervised and unsupervised techniques of data mining.	87%	Use of data mining strategies to detect fraudulent health insurance claims, focusing on pre-processing and correlation analysis.
6	Aayushi Verma	Propose an insurance claim fraud detection approach using data mining techniques, including association rule mining, clustering, and outlier detection methods.	80%	Integrate association rule mining, clustering, and outlier detection to identify potential fraud cases among insurance claims.
7	Krittika Dutta	Utilize numerous data mining regression algorithms to predict the cost of health insurance, including decision tree, random forest, polynomial regression, and linear regression.	88%	Apply data mining regression algorithms to predict health insurance costs and evaluate their performance using different metrics.

8	Jin-Su Park	Discuss the use of a healthcare claims database for clinical studies, highlighting the characteristics and structure of Korea's health insurance system and the advantages of using the database for research purposes.	78%	Provide information on the establishment and availability of Korea's healthcare claims database for research purposes.
9	Jin-Su Park	Utilize a healthcare claims database for clinical studies, focusing on the characteristics and structure of Korea's health insurance system.	88%	Provide an outline of the healthcare claims database, its structure, use in research, and the importance of the National Health Insurance Service (NHIS) and the Health Insurance Review Agency (HIRA) in managing the database.
10	Thomas George K	Proposing the use of machine learning strategies for detecting auto/vehicle insurance fraud, specifically focusing on decision tree, random forest, and naive Bayes algorithms.	90%	Utilize machine learning in establishing a predictive model to detect fraudulent auto/vehicle insurance claims, transforming raw data into feature vectors, and applying classification algorithms for fraud detection.
11	Lavanya Settipalli	Propose the Weighted MultiTree (WMT) approach for provider profiling and labeling of deceptive health insurance claims using a Directed Acyclic Graph (DAG) called MultiTree and assigning weights based on relative frequency.	73%	Develop a method for provider profiling and labeling of deceptive health insurance claims using the Weighted MultiTree (WMT) approach, aiming to enhance fraud detection accuracy by assigning weights and utilizing provider details.
12	Sharifa Rigga Mambo	Utilize classification models including Naïve Bayes, Decision tree, and K-Nearest Neighbor for detection of fraudulent health insurance claims, primarily focusing on the performance of the Naïve Bayes model.	91%	Utilize classification algorithms for fraudulent detection in health insurance claims, with primary focus on the Naïve Bayes model, aiming to identify potential fraud and develop a prototype for real-time fraud detection.
13	Shamitha S.K, V Ilango	Proposing a framework of fraud detection in health insurance claims using artificial neural networks, specifically the Multi-Layer Perceptron (MLP) with genetic algorithm optimization and Principal Component Analysis (PCA) for feature selection.	90%	Develop a framework of fraudulent detection in health insurance claims using artificial neural networks, genetic algorithm optimization, and PCA for feature selection, aiming to improve efficiency and accuracy in fraud detection.
14	P Naga Jyothi	Present the Supervised Outlier Detection Approach in healthcare Claims (SODAC) method, combining data preprocessing with outlier detection using statistical and distance-based approaches.	80%	Develop a method for outlier detection in healthcare claims using SODAC approach, incorporating statistical and distance-based techniques to identify potential fraudulent claims in large-scale datasets.
15	Md Enamul Haque	Propose the fraudulent detection approach for health insurance claims using the Mixture of Clinical Concepts (MCC) model, representing claims as mixtures of latent clinical concepts and enhancing detection with LSTM and RPCA strategies.	82%	Develop an approach of fraudulent detection in health insurance claims using MCC model, representing claims as combination of clinical concepts and enhancing detection with LSTM and RPCA extensions to improve fraud detection capabilities.
16	Vimala Balakrisnan	Utilization of systematic literature review (SLR) approach to analyze machine learning algorithms, features, and performance measures used in cyberbullying detection, centered on the challenges and future expansions in automatic cyberbullying detection models.	89%	Conducting of systematic literature review (SLR) to examine machine learning algorithms, features, and performance measures in cyberbullying detection, aiming to identify challenges and future research expansios in the field.
17	Shashank Agarwal	Present an intelligent machine learning strategy for fraud detection in medical claim insurance, incorporating domain knowledge, expert rules, and the K-means clustering algorithm to distinguish between legitimate and potentially fraudulent claims.	93%	Develop an intelligent machine learning approach for fraud detection in medical claim insurance, utilizing domain knowledge, expert rules, and K-means clustering algorithm to enhance accuracy and efficiency in fraud detection.

V. CONCLUSION:

Machine learning provides a powerful tool for detecting fraudulent health insurance claims. Insurance companies can create robust models to detect suspicious activity by combining various supervised learning algorithms such as logistic regression, decision trees, random forests, support vector machines, and gradient boosting, as well as appropriate preprocessing techniques and feature extraction methods. These algorithms may examine previous claims data, identify important traits, and forecast the possibility of fraud in real time, allowing for early action and risk mitigation. Additionally, ensemble methods, anomaly detection techniques, and deep learning approaches can improve detection accuracy and responsiveness to changing fraud patterns. Finally, by using effective machine learning technologies, insurers may considerably decrease financial losses, protect their customers from fraudulent activities, and maintain trust and integrity.

VI. References:

[1] : L. Ismail and S. Zeadally, "Healthcare Insurance Frauds: Taxonomy and Blockchain-Based Detection Framework (Block-HI)", IT Prof., vol. 23, no. 4, pp. 36-43, 2021.
[2] :"The challenge of health care fraud", Jan. 2020, [online] Available: https://www.nhcaa.org/resources/health-care anti-fraud resources/the-challenge-of-health-care fraud.aspx
[3] : B. Xu, H. Shen, B. Sun, R. An, Q. Cao and X. Cheng, "Towards consumer loan fraud detection: Graph neural networks with role-constrained conditional random field", Proc. AAAI Conf. Artif. Intell., pp. 4537-4545, 2021.
[4] : Z. Niu, D. Cheng, L. Zhang and J. Zhang, "Visual analytics for networked-guarantee loans risk management", Proc. Pacific Visual. Symp., pp. 160-169, 2018.
[5] : M. Bärtl and S. Krummaker, "Prediction of claims inexport credit finance: a comparison of four machine learning techniques", Risks, vol. 8, no. 1, 2020.
[6] : K. Amer, ‘‘The healthcare juggernaut,’’ Dawn news, Karachi, Pakistan, Tech. Rep., 2018.
[7] : C. Zhang, X. Xiao, and C. Wu, ‘‘Medical fraud and abuse detection system based on machine learning,’’ Int. J. Environ. Res. Public Health, vol. 17, no. 19, p. 7265, Oct. 2020.
[8] : C. Verma, V. Stoffova, Z. Illes, S. Tanwar, and N. Kumar, ‘‘Machine learning-based student’s native place identification for real time,’’ IEEE Access, vol. 8, pp. 130840–130854, 2020.
[9] : L. Zhang, J. Lin, and R. Karim, “Adaptive kernel density-based anomaly detection for nonlinear systems,” Knowledge Based Systems, vol. 139, pp. 50–63, 2018.
[10] : I. Kose, M. Gokturk, and K. Kilic, “An interactive machine-learning based electronic fraud and abuse detection system in healthcare insurance,” Applied Soft Computing, vol. 36, pp. 283–299, 2015.
[11] : R. A. Bauder and T. M. Khoshgoftaar, “A probabilistic programming approach for outlier detection in healthcare claims,” in Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on. IEEE, 2016, pp. 347–354.
[12] : J. Wang and S. Luo, “Augmented beta rectangular regression models: A bayesian perspective,” Biometrical Journal, vol. 58, no. 1, pp. 206–221, 2016.
[13] : Centers for Medicare and Medicaid Services, “ICD- 10,” https://www.cms.gov/Medicare/Coding/ICD10/, 2020, accessed January, 2020.
[14] : Medical Billing and Coding, “HCPCS codes,” https://www.medicalbillingandcoding.org/hcpcs-codes/, 2020, accessed January, 2020.
[15] : American Academy of Professional Coders, “CPT codes,” https://coder.aapc.com/cpt-codes, 2020, accessed January, 2020. [16] : S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735– 1780, 1997.
[17] : D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.
[18] : T. Hofmann, “Probabilistic latent semantic analysis,” in Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.
[19] : E. J. Candes, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” Journal of the ACM (JACM), vol. 58, no. 3, p. 11, 2011.
[20] : Centers for Medicare and Medicaid Services, “Research, statistics, data and systems,” https://www.cms.gov/Research-Statistics-Data-and Systems/Downloadable-Public-Use-Files/SynPUFs/DE Syn PUF, 2020, accessed January, 2020.
[21] : R. A. Bauder and T. M. Khoshgoftaar, “The detection of medicare fraud using machine learning methods with excluded provider labels,” in The Thirty- First International Flairs Conference, 2018.