Classifying clinically actionable genetic mutations using KNN and SVM

CHIVUKULA, Rohit, TANGIRALA, Jaya Lakshmi <http://orcid.org/0000-0003-0183-4093>, UDAY, Sanku Satya and PAVANI, Satti Thanuja

Published version

CHIVUKULA, Rohit, TANGIRALA, Jaya Lakshmi, UDAY, Sanku Satya and PAVANI, Satti Thanuja (2021). Classifying clinically actionable genetic mutations using KNN and SVM. Indonesian Journal of Electrical Engineering and Computer Science, 24 (3), 1672-1679.

Indonesian Journal of Electrical Engineering and Computer Science
Vol. 24, No. 3, December 2021, pp. 1672~1679
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v24.i3.pp1672-1679

Classifying clinically actionable genetic mutations using KNN and SVM

Rohit Chivukula[1], T. Jaya Lakshmi[2], Sanku Satya Uday[3], Satti Thanuja Pavani[4]
[1]School of Computing and Engineering, University of Huddersfield, Huddersfield, United Kingdom
[2,3,4]Department of Computer Science and Engineering, SRM University, Andhra Pradesh, India

Article Info

Article history:

Received Feb 24, 2021
Revised Oct 19, 2021
Accepted Oct 27, 2021

Keywords:

Genetic mutation
KNN
Log-loss
Multi-class classification
Personalized cancer treatment
Problem
SVM

ABSTRACT

Cancer is one of the major causes of death in humans. Early diagnosis of genetic mutations that cause cancer tumor growth leads to personalized medicine to the decease and can save the life of majority of patients. With this aim, Kaggle has conducted a competition to classify clinically actionable gene mutations based on clinical evidence and some other features related to gene mutations. The dataset contains 3321 training data points that can be classified into 9 classes. In this work, an attempt is made to classify these data points using K-nearest neighbors (KNN) and linear support vector machines (SVM) in a multi class environment. As the features are categorical, one hot encoding as well as response coding are applied to make them suitable to the classifiers. The prediction performance is evaluated using log loss and KNN has performed better with a log loss value of 1.10 compared to that of SVM 1.24.

Corresponding Author:

T. Jaya Lakshmi
Department of Computer Science and Engineering, SRM University
Andhra Pradesh, India
Email: jaya.phd.hcu@gmail.com

1. INTRODUCTION

Cancer is one of the leading causes of death in humans. The main cause of cancer is gene mutations within cells. The survival and recovery of a cancer patient highly depends on diagnosing and treating it at an early stage. Personalized treatment for a cancer patient can be designed if the mutations are understood in advance [1]. Though there are advanced treatments for cancer, the understanding of genetic mutations is limited by a large amount of manual work [2]. Memorial Sloan Kettering Cancer (MSKCC) has come up with an expert knowledge base describing the annotations of nine classes of clinically actionable genes. This problem can be modelled as a multi-class classification problem. In the year 2017, MSKCC launched a competition, on Kaggle [3] to facilitate personalized cancer treatment to the patients [4]. The task is to develop classification models utilizing the text in the given medical articles that can predict oncogenicity and mutation effect of the genes specified in the content. In this assessment, it is proposed to understand the data and to examine two classification algorithms on the dataset.

2. PROBLEM STATEMENT

2.1. Problem statement

Given a set of genetic mutations with their associated features, one feature being a long text representing the clinical evidence about the mutation given by an expert, and labelled with one of nine possible class labels, the problem of gene mutation classification is to predict a class label from 1 to 9 for a gene

mutation instance with missing label. As the prediction is not binary, but among 1 to 9, this is a multi-class classification problem.

2.2. Dataset description

The competition launcher [3] provided a separate train and test sets consisting of the genetic information of each instance is spanned over two different files as:

- training_variants (ID, gene, variations, and class): Contains the information about the genetic mutations such as gene, variation and class to which this mutation belongs to.
- training_text (ID, text): Contains a long text describing the clinical evidence that human experts use to classify the genetic mutations.
- Both these data files have a common column called ID through which they can be joined.

Similarly, two files for testing with similar information are provided. Genetic mutations are classified into nine different classes. The competition launcher provided a separate train and test sets. An example data point in the training variants is shown in Table 1. There are 3,321 training instances and 5,667 testing instances in the dataset.

Table 1. Example data point

File	Example data point
training_variants (ID, Gene, Variation, Class)	0, FAM58A, Truncating Mutations,1
training_text (ID,Text)	0||<long text describing abstract of clinical >

2.3. Exploratory data analysis

Performing exploratory data analysis on dataset helps in understanding the data in depth. The distribution of instances over 9 classes in the training set is shown in Figure 1. From the histogram, Class 7 has highest number of instances containing 28.6% of total training dataset and class 8 with lowest number (0.56%). For a classification task, exploratory data analysis (EDA) further helps in determining which features are useful for the machine learning task.

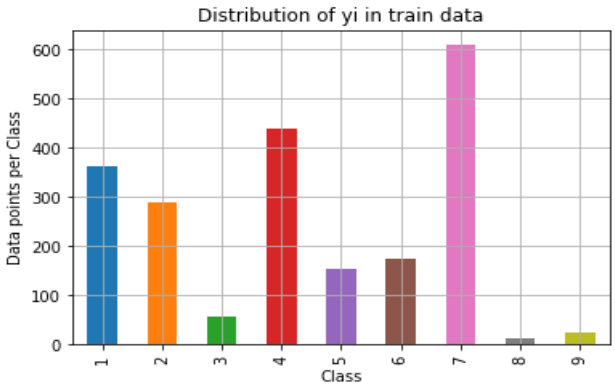

Figure 1. Distribution of instances over 9 classes

2.4. Evaluation measures

The following evaluation metrics are used in this work.

- Log loss [5], [6]: Log loss metric is commonly used in multi-class classification problems. This metric considers the probabilities of each problem instance belonging to each class and sums up. As the probabilities are closure to the true class membership, the better will be the value of log loss. It is a measure of uncertainty, so a low log loss means a low uncertainty. The ideal value of log loss is 0 for a strong classifier.

In the gene mutation dataset, for each ID in the test set, the classification algorithm must predict a probability for each of the 9 classes a genetic mutation can be classified on. Many of the classification problems evaluate their performance using accuracy. Accuracy can be sometime misleading if the dataset is imbalanced. Confusion matrix gives a better picture where the chosen classification model is performing well and what types of errors it is making.

- Confusion matrix: A confusion matrix for a n-class problem will be an n X n matrix, where columns correspond to the predicted class labels and the rows corresponds to the actual [7]-[9]. The main diagonal gives the correct predictions. That is, the cases where the actual values and the model predictions are the same. In this problem, the matrix is of size 9 X 9. Each cell [i,j] represents number of points of class i are predicted to belong to class j. The ideal value of confusion matrix C can be

$$C[i,j] = 0 \quad \text{if } i \neq j$$
$$= \quad \text{Number of instances of class i(or j)} \quad \text{if } i=j$$

- Precision matrix: Precision is the fraction of correctly predicted instances out of total predictions for a given class [8], [9]. Precision is good if cost of wrong belongingness prediction to a class.
- Recall matrix: Recall is the capture of correct predictions among total instances belonging to the class [8], [9]. Recall is good if cost of identifying an instance which is a member of the class. If a patient who is cancerous is not predicted, it is a huge loss to the patient.

3. RELATED LITERATURE

Machine learning can aid cancer diagnosis efficiently [10]. Clinical research data related to genetic detail can provide important insights on cancer [11]. Different types of cancers have been classified utilizing the gene expressions in [12]. The articles on biomedical data act as a strong source for the classification of clinically actionable genetic mutations [13].

To utilize the knowledge base available in the form of documents at PubMed database, a competition has been launched at Kaggle in which a dataset of oncogenes, their related mutations with articles. The aim of the competition can be basically viewed as text classification, but it is more challenging than that. Zhang *et al*., are the runners of the competition and have documented their solution in [14]. The primary problem they have faced in classification is that two genes can have same clinical evidence, with different mutations (class label). Therefore, the authors of [14] felt that text classification alone cannot solve the problem. Zhang *et al*., performed a novel feature engineering for obtaining three types of features. First being document features derived from clinical evidence documents; second the entity features are derived from mutations and clinical evidence and finally third are name features extracted from mutation, evidence using word embedding model. The authors of [15] apply one hot encoding on genes and their mutations to convert the features to numeric and TF-IDF mechanism to extract features from clinical evidence. Gangmin *et al*., [16] use TF-IDF technique to extract text features from clinical evidence data.

The work at [17] uses word embedding techniques and train their model using convolutional neural network (CNN) algorithm to classify cancer literature based on cancer hallmarks. Deep learning algorithms seem to work efficient in this domain. A comprehensive review of deep learning techniques is given in [18]. Errors in such sensitive problems can be very costly. An interesting cost sensitive approach for multi-class problem has been discussed in [19], which penalizes the misclassifications to improve learning in the model. Therefore, there is a strong need of more efficient models in future.

Reference [20] and [21] discuss the usage of metaheuristic algorithms after fuzzy modelling the problem. Abasi et al develop a cluster-based approach for text documents by treating text document clustering as discrete optimization problem [22]. Rustam et al propose a new feature selection method and use K-means clustering as the classifier using radial basis function and polynomial kernel function [23]. K-means clustering is sensitive to the parameter *k*. Sudha et al propose a novel modelling of clustering to get optimum number of clusters [24].

4. PROPOSED APPROACH

Any machine learning task, the approach in this work also follows steps such as performing exploratory data analysis, preprocessing, then training the classification model with train data and then evaluating the performance using performance metric. But feature engineering, preprocessing and classification parameters vary for different datasets. The approach is summarized in Figure 2.

4.1. Pre-processing

The machine learning algorithms that are intended to use in this work require the input and output variables as numeric. Therefore, the categorical features must be encoded to numbers. Two encodings are used in this work: one hot encoding [25] and response coding [26]. In one hot encoding, the feature is represented as a vector of size *n* if there are n distinct number of categories, with a 1/0 showing presence or absence of that value. Using response coding, a data point is represented as a vector showing the probability

of the data point belonging to a specific class given a category. For a *k*-class classification problem, there will be *k* new features which embed the probability of the datapoint belonging to each class based on the value of categorical data.

Gene and variation features are one hot encoded and clinical evidence feature is encoded using response coding. Gene and variation features are encoded using one hot encoding to make then suitable for machine learning task. Clinical evidence features, which is a long text contains several stop words, special characters and spaces. In the first step, all these are removed. Later, a vector of distinct words is built along with the count of each word in the instance. Retained only those words which occur more than 3 times and filtered all others. Then created two separate encoded vectors, one using one hot encoding and the other with response coding. Then the vectors are normalized. After the two types of encodings, the number of features is shown in the Table 2.

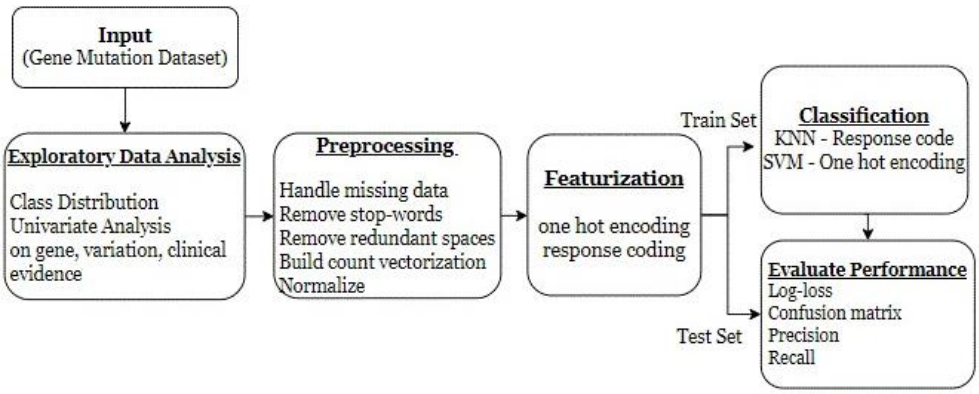

Figure 2. Proposed approach

Table 2. Total features after encoding

Encoding	Gene	Variation	Clinical Evidence	Total no. of features
One Hot encoding	229	1,960	54,850	57,029
Response coding	9	9	9	27

4.2. Machine learning algorithms

- Random model: In a random Model, we generate the NINE class probabilities randomly such that they sum to 1 for each test data point.
- K-nearest neighbors [27]: Based on the neighbors of the test data point, based on some distance measure. The test data point is assigned to the most common class among the K-nearest neighbors. Picking the value of k is challenge in this algorithm. Hyper parameter tuning will be performed by many which is to test various values for k and fix the best one that gives better prediction accuracy [28]. In this work, different values ranging from 5 to 100 have been experimented, the better log loss value has been obtained for k=15. The hyper parameter tuning has been shown n Figure 3. K-nearest neighbors (KNN) is local method and model need not be built in advance. But it is less susceptible for noise.
- Linear support vector machines (SVM) [29]: A support vector machine for binary classification problem is based on the idea of finding a hyper plane that nest separated the data points belonging to two classes. In a multi class classification problem with n classes, a one vs one approach is used. In this approach, n*(n-1)/2 number of binary classifiers are built for each pair of classes. The target class of the test data point is chosen by majority voting. The regularization parameter is tuned experimented between 0.00001 and 100 and obtained a lower log loss for 0.01. Therefore, a regularization parameter of 0.01 is used in final experimentation. The regularization parameter tuning [30] is shown in Figure 4.

After training a classification model on training data, the performance is to be evaluated on a test set. As it is clearly seen from Figure 1 that class distribution is not balanced. Therefore, stratified sampling is performed on training set to form test as well as cross validation sets to maintain the similar distribution.

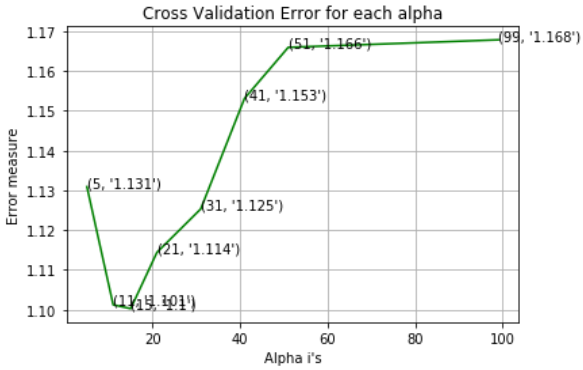

Figure 3. Hyper parameter tuning for KNN

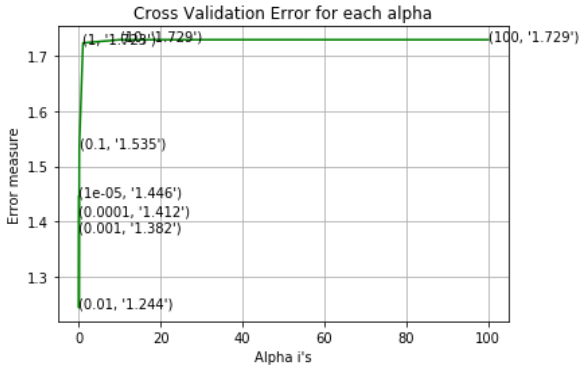

Figure 4. Hyper parameter tuning for SVM

5. RESULTS

The classification results of log loss evaluation metric for gene mutation dataset are tabulated in Table 3. It can be observed that KNN has shown more prediction accuracy compared to linear SVM in terms of log loss. As already have mentioned, the less the log loss value, better the prediction. The performance of KNN has been doubled compared to random prediction and increased slightly than performance of SVM.

KNN is lazy classifier, and no model needs to be built in advance. It is a local method and works based on nearest neighbors. The training time for SVM is high. The confusion matrix, precision and recall matrices are given in Figures 5 to 7. The diagonal value corresponding to a class in confusion matrix gives the idea about the prediction. The diagonal entry, precision and recall for class 7, 9, 3 and 8 of KNN classifier are given in Table 4.

Table 3. Logg loss results

S.No	Classifier	Log Loss	Number of misclassified points
1	Random	2.53	2.50
2	KNN	1.10	0.39
3	SVM	1.24	0.39

Table 4. Analysis of class 7, 9, 3 and 8 obtained using KNN

Class	% of Test data points	Diagonal value in Confusion matrix	Precision	Recall
7	28.7%	119	0.68	0.77
9	1%	3	1	0.5
3	2.7%	1	0	0
8	0.6%	0	0	0

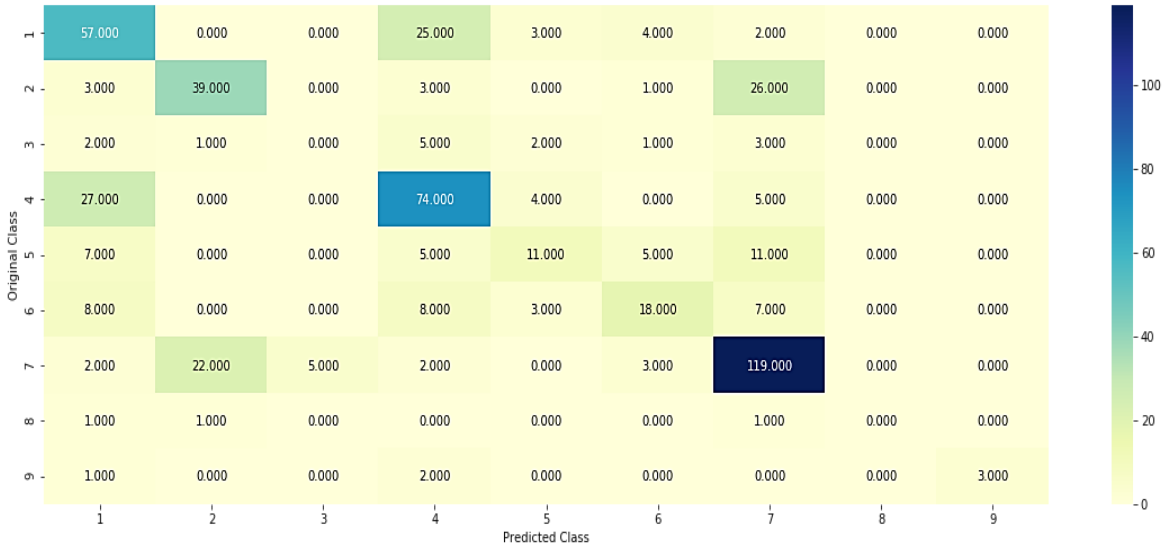

Figure 5. Confusion matrix of KNN classifier

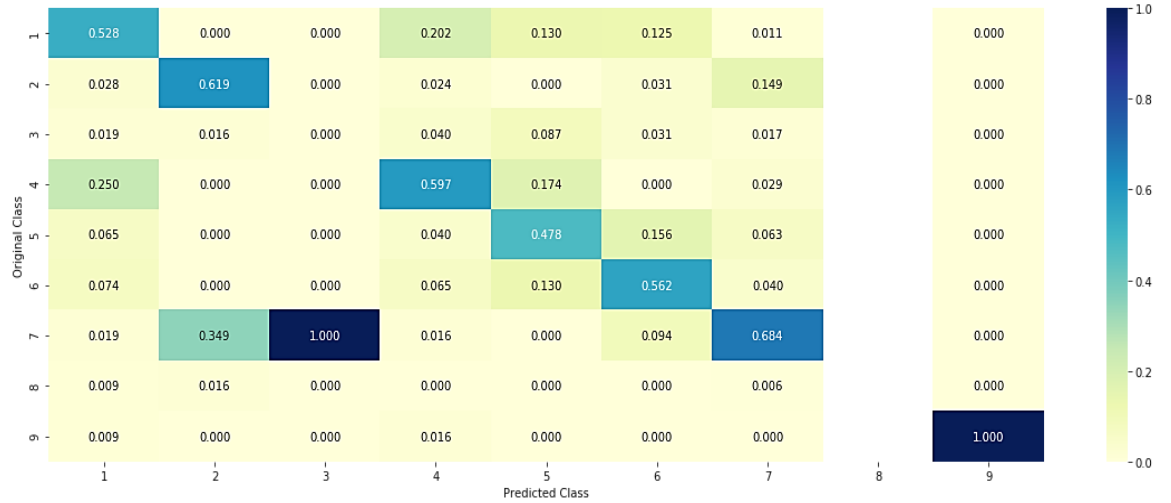

Figure 6. Precision matrix of KNN

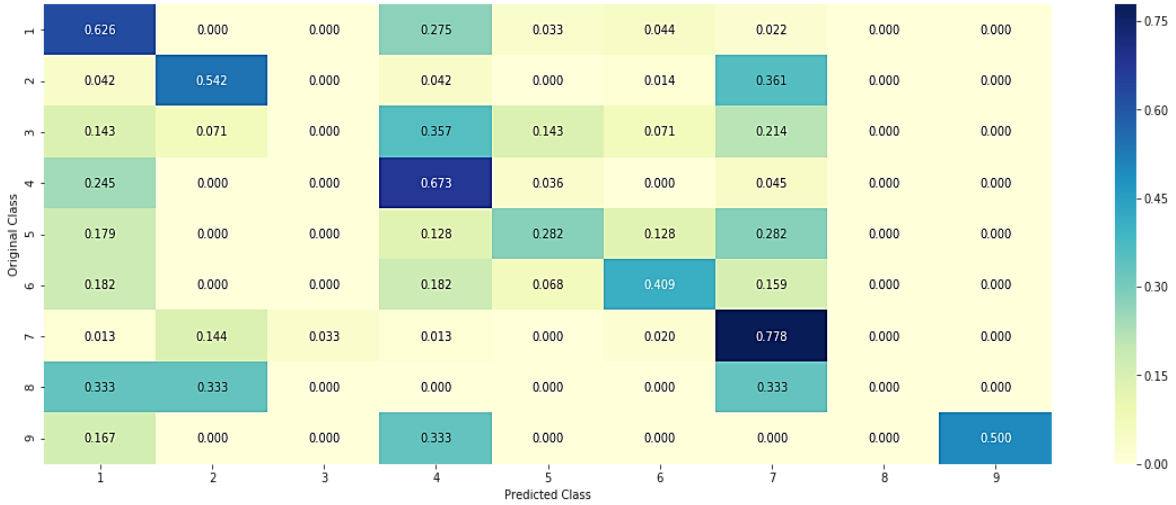

Figure 7. Recall matrix of KNN

In the dataset, class 7 is major class with 28% of data points. KNN classifier has acquired a precision value of 0.68 and recall of 0.77 for class 7 indicating that 68% of the data points in the predicted and 77% in actual points correctly. The two classifiers KNN as well as SVM failed in predicting the instances of class 3 and 8 with zero precision and recall. Class 3 and 8 have least number of data points in the dataset. I guess this is because the standard classifiers are designed for balanced datasets. Further investigation is needed for analyzing the prediction accuracy for class 3 and 8. Comping to class 9 containing 1% of test instances, KNN could obtain an ideal precision of 1 and recall of 0.5. This means, whatever the instances that KNN classified into class 9 are accurate, but it could predict only half the existing instances of class 9. This observation indicates that precision or recall alone cannot be taken into consideration while analyzing.

6. CONCLUSION AND FUTURE WORK

A personalized treatment for cancer can be efficiently designed if the medical experts have the pre information of genetic mutations. Machine learning methods can be effectively used to classify the genetic mutations based on clinically actionable data. In this work, K-nearest neighbors and linear support vector machines are applied on the public dataset available at Kaggle competition. KNN is found to achieve a better classification accuracy over Linear SVM in terms of log loss. In future, it is intended to use the efficient text extraction features like TF-IDF and doc2vec and train the models using ensemble framework. As this

problem is imbalanced, class balancing mechanisms such as under sampling, over sampling can be used to obtain more accurate results. Deep learning methods seem to work efficient on text processing. In future, we would work on deep learning techniques on the dataset.

REFERENCES

[1] L. Chin, J. N. Andersen, and P. A. Futreal, "Cancer genomics: from discovery science to personalized medicine," *Nature medicine*, vol. 17, no. 3, pp. 297-303, 2011, doi: 10.1038/nm.2323.

[2] T. Cheng and X. Zhan, "Pattern recognition for predictive, preventive, and personalized medicine in cancer," *EPMA Journal*, vol. 8, no. 1, pp. 51-60, 2017, doi: 10.1007/s13167-017-0083-9.

[3] Kaggle, Personalized Medicine: Redefining Cancer Treatment, 2018. [Online]. Available: https://www.kaggle.com/c/msk-redefining-cancer-treatment/overviewsd

[4] NIPS 2017 Competition Track, 2017. [Online]. Available: https://nips.cc/Conferences/2017/CompetitionTrack

[5] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27-38, 2009, doi: 10.1016/j.patrec.2008.08.010.

[6] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85 no. 3, pp. 333-359, 2011, doi: 10.1007/s10994-011-5256-5.

[7] J. T. Townsend, "Theoretical analysis of an alphabetic confusion matrix," *Perception & Psychophysics*, vol. 9, no. 1, pp. 40-50, 1971, doi: 10.3758/BF03213026.

[8] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[9] K. M. Ting, "Confusion Matrix," *Encyclopaedia of Machine Learning and Data Mining*, pp. 260, Boston, USA: Springer, 2017, doi: 10.1007/978-1-4899-7687-1_50.

[10] J. Goecks, V. Jalili, L. M. Heiser, and J. W. Gray, "How machine learning will transform biomedicine," *Cell*, vol. 181, no. 1, pp. 92-101, 2020, doi: 10.1016/j.cell.2020.03.022.

[11] T. C. Carter and M. M. He, "Challenges of identifying clinically actionable genetic variants for precision medicine," *Journal of healthcare engineering*, vol. 2016, 2016, doi: 10.1155/2016/3617572.

[12] G. C. Cawley and N. L. C. Talbot, "Gene selection in cancer classification using sparse logistic regression with Bayesian regularization," *Bioinformatics*, vol. 22, no. 19, pp. 2348-2355, 2006, doi: 10.1093/bioinformatics/btl386.

[13] N. Peng, H. Poon, C. Quirk, K. Toutanova, and W. Yih, "Cross-sentence n-ary relation extraction with graph lstms," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 101-115, 2017, doi: 10.1162/tacl_a_00049.

[14] X. Zhang, *et al*., "Multi-view ensemble classification for clinically actionable genetic mutations," *The NIPS'17 Competition: Building Intelligent Systems*, pp. 79-99, 2018, doi: 10.1007/978-3-319-94042-7_5.

[15] R. N. Waykole and A. D. Thakare, "Intelligent Classification of Clinically Actionable Genetic Mutations Based on Clinical Evidences," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697395.

[16] G. Li and B. Yao, "Classification of Genetic Mutations for Cancer Treatment with Machine Learning Approaches," *International Journal of Design, Analysis and Tools for Integrated Circuits and Systems*, vol. 7, no. 1, pp. 63-67, 2018.

[17] N. Ali, A. H. AbuEl-Atta, and H. H. Zayed, "Enhancing the performance of cancer text classification model based on cancer hallmarks," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 2, pp. 316-323, 2021, doi: 10.11591/ijai.v10.i2.pp316-323.

[18] C. Mishra and D. L. Gupta, "Deep machine learning and neural networks: An overview," *IAES International Journal of Artificial Intelligence*, vol. 6, no. 2, pp. 66-73, 2017, doi: 10.11591/ijai.v6.i2.pp66-73.

[19] M. A. Febriantono, S. H. Pramono, Rahmadwati, and G. Naghdy, "Classification of multiclass imbalanced data using cost-sensitive decision tree C5. 0," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 1, pp. 65-72, 2020, doi: 10.11591/ijai.v9.i1.pp65-72.

[20] M. S. Nordin et al., "A comparative analysis of metaheuristic algorithms in fuzzy modelling for phishing attack detection," *Indonesian Journal of Electrical and Computer Engineering*, vol. 23, no. 2, pp. 1146-1158, 2021, doi: 10.11591/ijeecs.v23.i2.pp1146-1158.

[21] S. S. M. Ali, A. H. Alsaeedi, D. Al-Shammary, H. H. Alsaeedi, and H. W. Abid, "Efficient intelligent system for diagnosis pneumonia (SARS-COVID19) in X-ray images empowered with initial clustering," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 1, pp. 241-251, 2021, doi: 10.11591/ijeecs.v22.i1.pp241-251.

[22] A. K. Abasi, A. T. Khader, M. A. Al-Betar, S. Naim, M. A. Awadallah, and O. A. Alomari, "Text documents clustering using modified multi-verse optimizer". *International Journal of Electrical and Computer Engineering*, vol. 10, no. 6, pp. 6361-6369, 2020, doi: 10.11591/ijece.v10i6.pp6361-6369.

[23] Z. Rustam and S. Hartini, "New feature selection based on kernel," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 4, pp. 1569-1577, 2020, doi: 10.11591/eei.v9i4.1959.

[24] V. Sudha and H. A. Girijamma, "Novel modelling of clustering for enhanced classification performance on gene expression data," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 2, pp. 2088-8708, 2020, doi: 10.11591/ijece.v10i2.pp2060-2068.

[25] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied Multiple Regression*. New York, USA: Routledge, 2013, doi: 10.4324/9780203774441.

[26] K. Wiegand and E. Wascher, "Response coding in the Simon task," *Psychological Research*, vol. 71, no. 4, pp. 401-410, 2007, doi: 10.1007/s00426-005-0027-1.
[27] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. USA: Springer, 2009.
[28] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for knn classification," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 3, pp. 1-19, 2017, doi: 10.1145/2990508.
[29] D. A. Pisner and D. M. Schnyer, "Support vector machine," *Machine Learning*, pp 101-121, Cambridge, UK: Academic Press, 2020, doi: 10.1016/B978-0-12-815739-8.00006-7.
[30] L. Wang, *Support vector machines: theory and applications*, USA: Springer, 2005.

BIOGRAPHIES OF AUTHORS

Rohit Chivukula is Masters student in Computing with University of Huddersfield, Huddersfield, United Kingdom. His research interests include Machine Learning, Data Analytics and Natural Language Computing He has acquired his bachelor's degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India. He can be contacted at email: rohit.chivukula.kingdom@gmail.com.

T. Jaya Lakshmi is working as an Assistant Professor in the Department of Computer Science and Engineering, SRM University, Amaravathi, AP, India. She received her Ph.D. Degree for her work on "Link Prediction in Heterogeneous Social Networks" from the School of Computer and Information Sciences, University of Hyderabad, India in 2019. She is reviewer of reputed international journals. She has a overall teaching experience of 22 years. Her research interests include graph mining, recommender systems, Natural Language Processing and security analytics. She can be contacted at email: jaya.phd.hcu@gmail.com)

Sanku Satya Uday is Bachelors student in Computer Science and Engineering with SRM University., Andhra Pradesh, India. His research interests include Machine Learning, Data Analytics and Natural Language Computing. He can be contacted at email: sanku_satya@srmap.edu.in.

Satti Thanuja Pavani is Bachelors student in Computer Science and Engineering with SRM University, Andhra Pradesh, India. Her research interests include Machine Learning, Data Analytics and Natural Language Computing. She can be contacted at email: satti_thanuja@srmap.edu.in.