Oncogenic Genes Identification using NLP and Ensemble Learning Approach

Bharathi Mohan G
Department of Computer science and Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham, Chennai, India
g_bharathimohan@ch.amrita.edu

R. Prasanna Kumar
Department of Computer science and Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham, Chennai, India
kumarprasanna.r@gmail.com

Vaibhav Chauhan
School of Biotechnology,
Amrita Vishwa Vidyapeetham,
Amritapuri Campus, Kollam, India
vaibhavchauhan200493@gmail.com

Naman Chauhan
Department of Computer science and Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham, Chennai, India
namanchauhan200293@gmail.com

***Abstract*—Oncogenic gene identification plays a pivotal role in advancing cancer research and treatment strategies. Accurate prediction and classification of genetic mutations are essential for targeted interventions and personalized medicine. This research leverages advanced machine learning techniques, including Naive Bayes, Logistic Regression, Support Vector Machines, Random Forest Classifier, and Partial Least Square Regression, along with ensemble learning, to enhance oncogenic gene detection. The dataset containing the gene variations along with the proper clinical evidence for each variation underwent rigorous pre-processing to ensure uniformity, cleanliness and normalization using one-hot encoding and Word vectorization. Performance metrics such as F1 Score, precision, recall, and accuracy were employed to evaluate the models. The ensemble model emerged as the most effective, as by combining all the machine learning models it achieved an outstanding accuracy rate of 97.8%. This exceptional accuracy signifies a significant advancement over traditional methods and underscores the transformative potential of computational technologies in oncology. The findings pave the way for more precise and personalized cancer treatment regimens.**

***Index Terms*—Oncogenic genes, One-hot encoding, Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, Partial Least square regression, Ensemble Learning.**

I. Introduction

Oncogenic gene identification is a cornerstone in cancer research, offering critical insights into the genetic alterations that underlie tumorigenesis and cancer progression. Understanding these genetic mutations is crucial for the development of targeted therapies and personalized treatment approaches, which have the potential to revolutionize cancer care by improving patient outcomes and reducing adverse effects. As the field of oncogenomics continues to expand, there is an escalating demand for advanced computational techniques to analyze and interpret the vast amounts of genomic and clinical data generated. Despite the progress made in genomics and computational biology, the identification of oncogenic genes remains a complex and challenging task. Traditional methods often rely on manual curation and interpretation of genomic data, which can be time-consuming, subjective, and limited in scope. Moreover, the heterogeneity and complexity of cancer genomes require sophisticated analytical tools capable of detecting subtle genetic alterations and understanding their functional implications.

In recent years, there has been a significant advancement in oncogenic gene identification leveraging computational and bioinformatics approaches. Natural Language Processing techniques have been employed to extract valuable insights from scientific literature and clinical records. This method has facilitated the identification of genes associated with specific cancer types and provided a deeper understanding of their roles in oncogenesis. One major challenge has been the integration of heterogeneous data sources such as genomic sequences, clinical records, and scientific literature. Traditional methods often struggle to combine these diverse data types effectively, leading to incomplete or biased analyses. Another concern is scalability; as genomic data volumes continue to grow, many existing approaches are not equipped to handle large datasets efficiently, resulting in increased computational costs and longer analysis times. Moreover, while machine algorithms can achieve high accuracy in predicting oncogenic mutations, their decision-making processes can be opaque and difficult to interpret. This lack of transparency limits the clinical utility of these models, as clinicians may be hesitant to trust predictions without understanding the underlying rationale. Lastly, cancer is a dynamic and complex disease with genetic mutations evolving over time. Traditional methods often lack the flexibility to adapt to these changes, resulting in static models that quickly become outdated.

In the proposed research, an extensive dataset is taken which contains two types of data files, one consisting of the gene variations and the other containing the clinical evidence for each variation. The data is then pre-processed and one-hot encoding is applied along with Count Vectorizer to transform data into numerical form that can be processed more robustly. Several algorithms like Support vector machine, Logistic Regression, Random Forest, Partial Least square regression and naive Bayes are trained and tested on the dataset and finally, all the models are stacked together to form an ensemble learning model that provides a more precise and accurate result.

The structure of this research study is as follows: Section 1 presents the research, Section 2 details the preliminary work needed for the study, Section 3 delves deeper into the proposed approach, Section 4 reports on the work's results, Section 5 talks about the outcomes of the proposed research, and Section 6 wraps up by talking about future endeavours.

II. Related Works

Using the first meeting documents from their doctor, Jose Nunez et al. [1] used NLP methods to estimate the survival of people with cancer, both in the short and long term. Results demonstrated high accuracy rates, comparable to or surpassing previous methods, suggesting the feasibility of predicting survival outcomes across various cancer types using readily available data without specific cancer type training. Marta Lovino, et al. [2] present a deep learning method using Convolutional Neural Networks (CNNs) to predict the oncogenic potential of gene fusion transcripts in cancer. By analyzing raw protein sequences, it offers flexibility and achieves an accuracy of about 72%, increasing to 86% for high-confidence predictions. Hong Yang, et al. [3] combine big data from The Cancer Genome Atlas with bioinformatics methods to identify 389 genes in liver hepatocellular carcinoma. ”Pathways in cancer” emerge as a priority, with four genes (BIRC5, E2F1, CCNE1, CDKN2A) showing diagnostic potential, validated by Oncomine. This gene pool demonstrates high diagnostic accuracy (AUC: 0.990), with BIRC5 and CCNE1 indicating poor prognosis. Kavita Iyer et al. [4] summarize findings from a Natural Language Programming analysis of 350K+ immuno-oncology publications (2000-2022), identifying 300+ emerging concepts. It presents a ”Trend Landscape Map” categorizing concepts and revealing rapid growth in protein targets like TROP2, alongside substance data analysis indicating higher commercial interest in protein/peptide sequences for cancer immunotherapy. Meijian Guan et al. [5] used NLP and machine learning to analyze cancer patient progress notes, identifying genomic-related treatment changes. Recurrent neural networks, especially bidirectional LSTM, outperformed traditional algorithms, with pre-trained word embedding boosting accuracy and reducing training time, showcasing the efficacy of NLP and RNN-based text mining in clinical settings. Advanced NLP and knowledge engineering methods are used in the DeepPhe program developed by Guergana K. Savova et al. [6] to automate the acquisition of detailed phenotypic information from the electronic medical records of cancer patients. Assessed using data from patients with breast cancer, it provides essential computational phenotyping techniques that enhance high-throughput sequencing analysis for the progress of precision cancer treatment. Vira Sorin et al. [7] proposed digital health data in oncology, including electronic health records and medical literature, presents untapped potential. However, the predominance of free-text formats limits computer analysis, necessitating advancements in data standardization and natural language processing for optimal utilization. Abdul wahab et al. [8] 4mCNLP-Deep computational model employs word embedding and CNN algorithms to predict N4 sites on the genome, surpassing existing predictors with high accuracy and specificity. An online web server facilitates easy access to results for experimental researchers. Jia Xu, et al. [9] explores the integration of AI in cancer genomics, focusing on its role in transforming big data into actionable insights for precision medicine. It reviews current AI applications, highlights challenges such as data requirements and algorithm transparency, and emphasizes the importance of preparing healthcare stakeholders for digitized healthcare. In order to classify PubMed entries related to the penetrance or frequency of germline genetic alterations, Yujia Bao et al. [10] constructed SVM and CNN models, which produced good accuracy results of 88–89%. For better clinical decision support, these models provide useful tools to help researchers and clinicians stay up to date on the growing body of knowledge about the relationships between genes and cancer. Meijian Guan, et al. [11] outperformed conventional ML techniques by using NLP and RNN-based classifiers to detect genomic-related treatment changes in cancer patients' progress notes. The best RNN in terms of accuracy, precision, recall, and F1 score was LSTM_Bi. Already trained word embeddings increased LSTM accuracy by 3.4% and decreased training time by more than 60%.

III. Preliminaries

A. Dataset information

The research's dataset was obtained from Kaggle [17], and in order to improve its efficiency, it underwent a number of data pre-processing procedures (covered in B subsection). The data comprises four files:

1) Training Variants: Comma-separated file with fields: ID, Gene, Variation, Class.
2) Training Text: Double pipe delimited file with fields: ID, Text.
3) Test Variants: Comma-separated file with fields: ID, Gene, Variation.
4) Test Text: Double pipe delimited file with fields: ID, Text.

The training files link genetic mutations to clinical evidence, while the test files are used for evaluating the classification of genetic mutations.

B. Data Pre-processing

Preprocessing of text is a crucial step in natural language processing tasks. It involves several steps such as removing stopwords, replacing special characters with spaces, eliminating multiple spaces, and lowercasing the text. These procedures guarantee uniformity and cleanliness of the text data for subsequent analysis. Once preprocessed, the gene_variations and text data can be merged based on a unique identifier

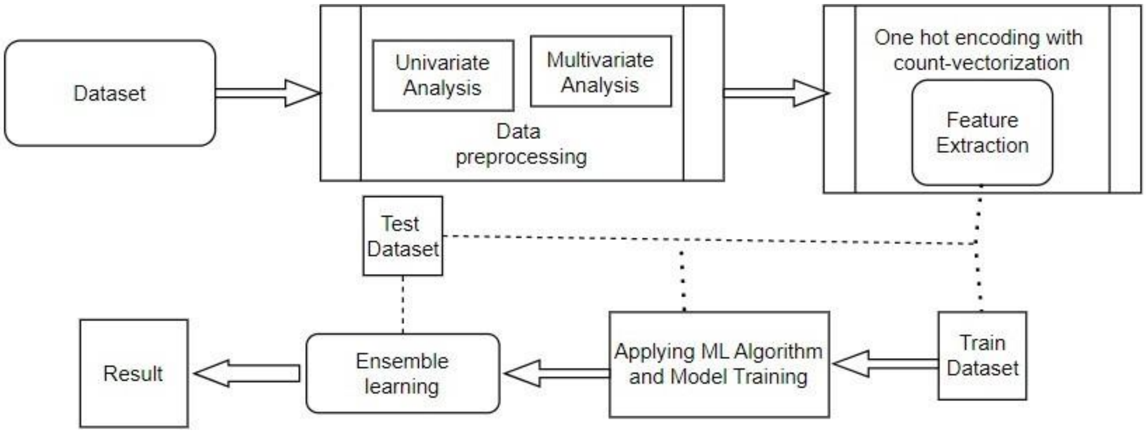


Fig. 1. Proposed Methodology

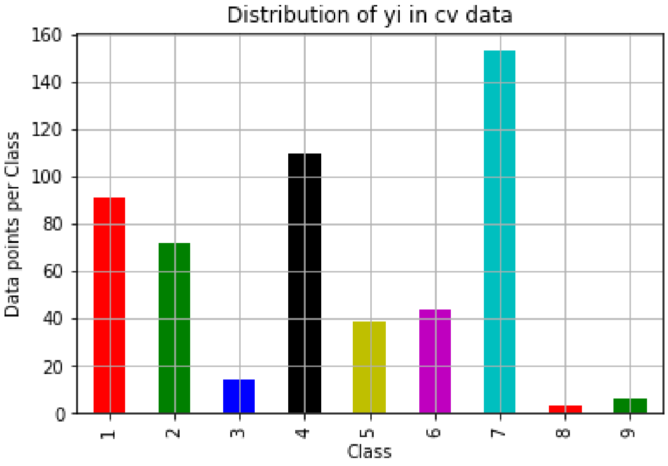


Fig. 2. Distribution of yi in Dataset

(ID), facilitating seamless integration for subsequent analysis. To save computational resources and time, the preprocessed dataframe can be stored for future use, eliminating the need to preprocess the data every time. Distribution analysis of the target variable (y_i), distribution for which can be seen in the figure 1 across train, test, and cross-validation datasets provides insights into the balance or skewness of the data, which is crucial for model training and evaluation. Univariate analysis involves examining individual features such as genes and variation. One-hot encoding of the gene feature using CountVectorizer converts categorical data into a numerical form suitable for ML frameworks. Similarly, one-hot encoding is applied to the variation feature. Moreover, analyzing the text feature involves determining the frequency of distinct words present in the train data, their frequencies, and featurizing the text field for further analysis.

The stability and usefulness of the text feature across different datasets (train, test, and cross-validation) are assessed. One-hot encoding[12] is applied to the text feature, followed by normalization of every feature to ensure consistency in scale. Checking the distribution of words helps in understanding the vocabulary and its relevance in predicting the target variable. Finally, a LR model is trained using only the text feature as input, and its performance across all datasets is evaluated to determine the stability and predictive power of the text feature.

IV. Methodology

The proposed work for Oncogenic gene detection begins with Data pre-processing of the dataset by applying various techniques like Univariate analysis and performing One-hot encoding on the data. Now, after converting the data into numerical form using one-hot encoding the ML models[13] described below are used to make insightful results on the given data and then the models are stacked to get the best result possible using ensemble learning.

A. *Naive Baiyes with one hot encoded feature*

Naive Bayes with one-hot encoded features[14], each categorical attribute is represented as binary features, enabling the class label-based algorithm's presumption of feature isolation. This encoding method is particularly suitable for Naive Bayes classification tasks, enabling efficient computation of conditional probabilities. Hyper-parameter tuning for Naive Bayes primarily revolves around parameters like Laplace smoothing for Gaussian Naive Bayes or smoothing parameters for categorical features in Multinomial Naive Bayes. Once the optimal hyper-parameters are determined through techniques like cross-validation, the model can be trained on the entire dataset with those settings. The model's ability to generalize can be evaluated by analyzing its efficacy on an independent test data. However, interpreting feature importance in Naive Bayes isn’t as direct as in other models due to its assumption of feature independence. Nonetheless, analyzing conditional probabilities associated with each feature given the class label can offer insights into their impact on classification decisions, both for correctly and incorrectly classified instances. This examination aids in understanding the model’s behaviour and identifying potential areas for improvement. The mathematical equation can be seen in equation 1.

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y) \qquad (1)$$

B. Logistic Regression with one-hot encoded features

Logistic Regression (LR) with one-hot encoded features involves transforming categorical attributes into binary features, a method that aligns well with LR’s assumption of a linear decision boundary. Class balancing techniques are essential to address imbalances in the dataset, ensuring the model does not become biased towards the majority class. Techniques such as oversampling, undersampling, or weighted loss functions can be applied to handle class imbalances effectively. Hyperparameter tuning is critical in LR to optimize parameters such as regularization strength, choice of solver, and regularization penalty. After tuning the hyperparameters, the LR model can be trained on the entire dataset with the optimized settings.The generalizing capacity of the model can then be tested by assessing its accuracy on an independent test dataset. The mathematical equation can be seen in equation 2.

$$p(X; b, w) = \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}} = \frac{1}{1 + e^{-w \cdot X + b}} \qquad (2)$$

C. Linear Support Vector Machines

Linear Support Vector Machines (SVMs)[15] are powerful classifiers commonly used for binary classification tasks. Optimization of variables like the parameter for regularization (C), which manages the trade-off between increasing the margin and minimizing the classification failure, and the loss function selection are examples of hyperparameter modification in SVMs. (e.g., hinge loss). Grid search or randomized search combined with cross-validation are typical approaches for hyperparameter tuning. Once the best hyperparameters are identified, the SVM model can be trained on the entire dataset using those settings. The framework’s adaptation capacity is then evaluated by analyzing its efficacy on an additional test data. While SVMs do not provide direct feature importance like some other algorithms, one can infer feature importance by analyzing the coefficients of the decision function. Features with larger coefficient magnitudes are considered more important in influencing the classification decision. The mathematical equation can be seen in equation 3.

$$\hat{y} = \begin{cases} 1 & : \; w^T x + b \geq 0 \\ 0 & : \; w^T x + b < 0 \end{cases} \qquad (3)$$

D. Random Forest Classifier

Random Forest Classifier with one-hot encoded features involves transforming categorical variables into binary features, which are then used to build decision trees within the ensemble. For optimal performance factors like the number of trees in the forest, the greatest depth of the trees, and the minimal number of samples needed to divide a node, random forest models require hyperparameter tweaking. To identify the ideal parameters, methods like grid searching and randomized search with cross-validation are frequently used. The Random Forest model can be learned on the complete dataset using these options when the ideal hyperparameters have been identified. The model’s capacity for generality can then be tested by analyzing the functionality on an independent test data. The mathematical equation can be seen in equation 4.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2 \qquad (4)$$

E. Partial Least Square Regression

Partial least squares regression (PLSR) is a tool we use in our research, on estimating yields. It’s a technique that can handle datasets with multiple interrelated factors. PLSR works by finding variables that capture the shared variation between predictor and response variables making it particularly useful in our investigation. By reducing dimensionality and addressing multicollinearity PLSR enhances the interpretability and reliability of our prediction model. Mathematically, it can be seen as Equation 5.

$$max_a corr^2(y, Xa) Var(Xa) \qquad (5)$$

Where, $||a||=1$, $\varphi_i^1 Sa=0$, $1=1, \ldots., m-1$

F. Stacking the Models

A potent ensemble learning method called stacking models[16] which can be seen in Figure 3, entails training several base models and aggregating the outcomes to get a final one. Hyperparameter tuning for stacked models typically involves tuning the hyperparameters of each base model individually, as well as determining the meta-model (or blender) hyperparameters, such as the choice of meta-learner and its parameters. Grid search or randomized search combined with cross-validation can be used for hyperparameter tuning in both base models and the meta-model. The meta-model is trained on the

TABLE I
PERFORMANCE MATRICES OF DIFFERENT MODELS

#	Algorithm	Recall	F1 Score	Precision	Accuracy
1	Naive Baiyes	0.9	0.91	0.922	92.6%
2	Logistic Regression	0.89	1.76	0.879	89.4%
3	SVM	0.86	1.548	0.842	86%
4	Random Forest	0.96	1.842	0.957	96.1%
5	PLSR	0.94	1.769	0.9375	94.3%

predictions made by the base models and learns to combine them optimally to make final predictions. After training, the stacked model can be tested on a separate test dataset to evaluate its performance. In maximum voting classification, predictions from multiple models are combined by taking the majority vote. The group with the highest number of votes is selected as the ultimate prediction, and every basic model in the piled ensemble throws a vote for the class predicted. Maximum voting classification is a simple yet effective way to combine predictions from multiple models and often leads to improved performance compared to individual models.

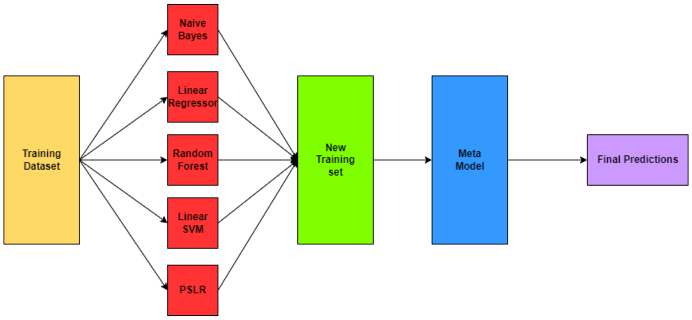


Fig. 3. Ensemble Learning Model

V. RESULTS AND DISCUSSION

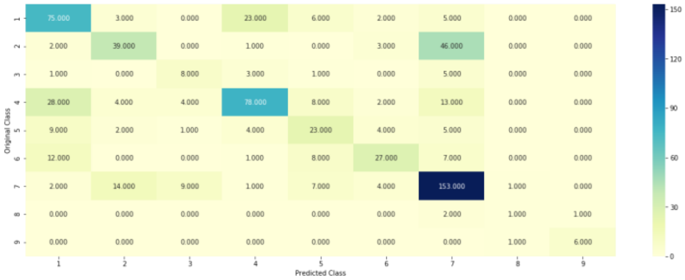
After applying Ensemble learning on all the models to get the best results out of it the models are compared based on several different matrices such as F1 Score, precision, Recall, accuracy and confusion matrix. Those are explained briefly below:

A. Confusion Matrix

In Figure 4, Confusion matrix can be seen for MLR, Decision Tree, SVM, Random Forest and PLSR, that gives us the TP, FN, FP, TN values essential for calculating other performance metrices like F1 score, precisiona and recall.

B. Performance Matrices

Using the different values obtained from the confusion matrix the other performance Matrices like F1 score, recall, precision and accuracy are calculated which can be seen in table 1.



	1	2	3	4	5	6	7	8	9
1	75.000	3.000	0.000	23.000	6.000	2.000	5.000	0.000	0.000
2	2.000	39.000	0.000	1.000	0.000	3.000	46.000	0.000	0.000
3	1.000	0.000	8.000	3.000	1.000	0.000	5.000	0.000	0.000
4	28.000	4.000	4.000	78.000	8.000	2.000	13.000	0.000	0.000
5	9.000	2.000	1.000	4.000	23.000	4.000	5.000	0.000	0.000
6	12.000	0.000	0.000	1.000	8.000	27.000	7.000	0.000	0.000
7	2.000	14.000	9.000	1.000	7.000	4.000	153.000	1.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	2.000	1.000	1.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	6.000

Fig. 4. Confusion Matrix

C. State-of-art Comparison with Existing work

The proposed model is compared with different existing models and the proposed model outperformed the existing model in terms of accuracy, which can be illustrated in figure 5.

VI. CONCLUSION AND FUTURE WORKS

Research work underscores the profound impact of integrating NLP and ML frameworks in oncogenic gene identification within cancer research. Through meticulous experimentation and evaluation, we have demonstrated the effectiveness of advanced computational methods, particularly ensemble learning, in achieving exceptional accuracy rates, with the ensemble model reaching 97.8%. These results signify a significant leap forward in the field, surpassing traditional approaches and highlighting the transformative potential of computational technologies. By accelerating the pace of cancer diagnosis and treatment strategies, this interdisciplinary approach holds promise for revolutionizing precision medicine, offering targeted interventions and personalized treatment regimens tailored to individual patients.

Moving forward, future research endeavors will focus on several key areas to further enhance the efficacy and applicability of NLP and ML techniques in cancer research. Firstly, efforts will be directed towards refining and optimizing existing algorithms to improve model interpretability and scalability. Additionally, exploring novel data sources and incorporating multi-modal data fusion techniques will broaden the scope of analysis, enabling a more comprehensive understanding of cancer biology. Furthermore, there is a pressing need to address challenges related to data standardization, privacy, and ethical considerations in the utilization of clinical data for research purposes. Collaborative efforts between interdisciplinary teams, including clinicians, bioinformaticians, and data

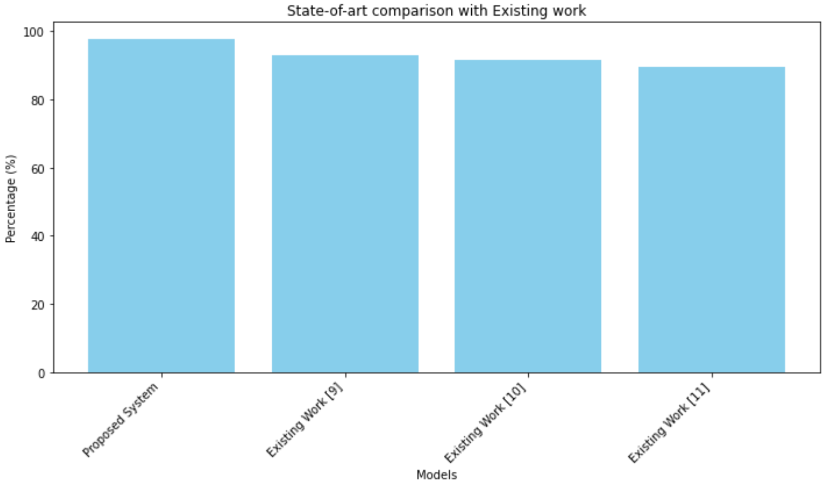


Fig. 5. State-of-art comparison with existing works

scientists, will be essential in driving forward the development and implementation of innovative computational approaches for advancing precision oncology.

REFERENCES

[1] Nunez J, Leung B, Ho C, Bates AT, Ng RT. Predicting the Survival of Patients With Cancer From Their Initial Oncology Consultation Document Using Natural Language Processing. JAMA Netw Open. 2023;6(2):e230813. doi:10.1001/jamanetworkopen.2023.0813

[2] Lovino, M.; Urgese, G.; Macii, E.; Di Cataldo, S.; Ficarra, E. A Deep Learning Approach to the Screening of Oncogenic Gene Fusions in Humans. Int. J. Mol. Sci. 2019, 20, 1645. https://doi.org/10.3390/ijms20071645

[3] Yang H, Zhang X, Cai X, Wen D, Ye Z, Liang L, Zhang L, Wang H, Chen G, Feng Z. 2017. From big data to diagnosis and prognosis: gene expression signatures in liver hepatocellular carcinoma. PeerJ 5:e3089 https://doi.org/10.7717/peerj.3089

[4] 1. Iyer K, Ivanov J, Tenchov R, Ralhan K, Rodriguez Y, Sasso J, et al. Emerging Targets and Therapeutics in Immuno-Oncology Landscape: Insights from Natural Language Processing Analysis . ChemRxiv. 2023; doi:10.26434/chemrxiv-2023-bpvfq.

[5] Meijian Guan, Samuel Cho, Robin Petro, Wei Zhang, Boris Pasche, Umit Topaloglu, Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes, JAMIA Open, Volume 2, Issue 1, April 2019, Pages 139–149, https://doi.org/10.1093/jamiaopen/ooy061

[6] Guergana K. Savova, Eugene Tseytlin, Sean Finan, Melissa Castine, Timothy Miller, Olga Medvedeva, David Harris, Harry Hochheiser, Chen Lin, Girish Chavan, Rebecca S. Jacobson; DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. Cancer Res 1 November 2017; 77 (21): e115–e118. https://doi.org/10.1158/0008-5472.CAN-17-0615

[7] Sorin, V., Barash, Y., Konen, E. and Klang, E., 2020. Deep-learning natural language processing for oncological applications. The Lancet Oncology, 21(12), pp.1553-1556.

[8] Wahab, A., Tayara, H., Xuan, Z. et al. DNA sequences performs as natural language processing by exploiting deep learning algorithm for the identification of N4-methylcytosine. Sci Rep 11, 212 (2021). https://doi.org/10.1038/s41598-020-80430-x

[9] Xu, J., Yang, P., Xue, S. et al. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. Hum Genet 138, 109–124 (2019). https://doi.org/10.1007/s00439-019-01970-5.

[10] Bao Y, Deng Z, Wang Y, Kim H, Armengol VD, Acevedo F, Ouardaoui N, Wang C, Parmigiani G, Barzilay R, Braun D, Hughes KS. Using Machine Learning and Natural Language Processing to Review and Classify the Medical Literature on Cancer Susceptibility Genes. JCO Clin Cancer Inform. 2019 Sep;3:1-9. doi: 10.1200/CCI.19.00042.

[11] Guan M, Cho S, Petro R, Zhang W, Pasche B, Topaloglu U. Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes. JAMIA Open. 2019 Apr;2(1):139-149. doi: 10.1093/jamiaopen/ooy061. Epub 2019 Jan 3.

[12] N. Chauhan, A. Kumar, G. V. Teja, B. M. G, P. K. R and Y. Kakarla, ”Deep Learning-Based Approach for Calorie Estimation in Indian Foods,” 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2024, pp. 828-833, doi: 10.1109/Confluence60223.2024.10463341.

[13] Mohan, G.B., Kumar, R.P., Elakkiya, R. and Gorantla, S., 2023, September. Enhancing Personality Classification through Textual Analysis: A Deep Learning Approach Utilizing MBTI and Social Media Data. In 2023 International Conference on Network, Multimedia and Information Technology (NMITCON) (pp. 01-06). IEEE.

[14] A. Kumar, Vidya H.A., Thejaswi A.H., “Transformer Fault classification using Na¨ıve Bayes (NB) algorithm,” at Journal of Emerging Technologies and Innovative Research, Vol 9, Issue 5, May 2022, pp-518-523.

[15] Akhil, V. M., K. J. Chandan, Deepa Itagi, and K. R. Prakash. ”Analyses of different methods of writing using SVM classifier.” In 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4), pp. 1-5. IEEE, 2021. https://doi.org/10.1109/C2I454156.2021.9689248

[16] G. Tallapureddy and D. Radha, ”Analysis of Ensemble of Machine Learning Algorithms for Detection of Parkinson’s Disease,” 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 354-361, doi: 10.1109/ICAAIC53929.2022.9793048

[17] https://www.kaggle.com/c/msk-redefining-cancer-treatment/data