IJITCE

ISSN 2347- 3657

International Journal of
Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

Deep CNN based Genomic variant classifier To predict Disease Susceptibility

Mohit Kumar Sharma [1] Annaboina Sujith Kumar [2], Kalangi Kranthi [3], S. Bavankumar [4]
[1,2,3]UG Scholar, Department of Computer Science and Engineering, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
[4]Assistant Professor, Department of Computer Science and Engineering, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
[1]smec.mohit@gmail.com, [2] sujithkumar132004@gmail.com, [3]kranthikal2002@gmail.com, [4]sbavankumar55@gmail.com

Abstract:

The rapid advancement in genomics has deepened our understanding of genetic variations and their link to human diseases. Genomic variants, including single nucleotide polymorphisms (SNPs), insertions, deletions, and complex rearrangements, influence disease susceptibility. Predicting disease risk based on these variants is crucial for personalized medicine. Historically, disease prediction relied on family history, linkage studies, and statistical models, which were slow, error-prone, and lacked accuracy. Before artificial intelligence, heuristic methods struggled with large genomic datasets, limiting predictive power. The growing need for precision medicine and the surge in genomic data demand automated, scalable solutions for accurate variant interpretation. Complex diseases such as cancer and cardiovascular disorders require advanced computational techniques. Traditional methods face challenges, including limited predictive accuracy, reliance on human expertise, and difficulty handling vast genomic data from next-generation sequencing. Addressing these issues requires robust, high-precision computational systems. This system utilizes deep convolutional neural networks (CNNs) to classify genomic variants and predict disease susceptibility with high accuracy. Deep learning automatically extracts features from large datasets, eliminating extensive manual feature engineering. By analyzing genomic sequences, it identifies disease-associated variants with improved precision, surpassing previous approaches. This AI-driven model enhances genomic variant analysis by reducing human intervention, efficiently processing large data, and delivering timely, accurate predictions. Its deep learning foundation revolutionizes disease risk assessment, paving the way for personalized medicine and early interventions. This breakthrough promises improved healthcare outcomes by enabling more precise, individualized treatment strategies.

Keywords: Genomic Variant, Random Forest, CNN, DNA, Genetic Acid, PCR, Sanger Sequencing.

1.INTRODUCTION

Genomic variants, including single nucleotide polymorphisms (SNPs), insertions, deletions, and complex rearrangements, play a crucial role in understanding genetic susceptibility to diseases. Identifying these variants accurately is essential for advancing personalized medicine, early disease detection, and targeted treatments. Traditional approaches, such as statistical models and manual interpretation, struggle with scalability and efficiency when handling vast genomic datasets. Deep convolutional neural networks (CNNs) offer a powerful solution by automatically extracting meaningful patterns from genomic sequences. This deep learning-based classifier enhances predictive accuracy, reduces human intervention, and processes large-scale genomic data efficiently, making it a vital tool for modern genomic research and precision medicine advancements.

Traditional disease prediction methods, including statistical models and manual analysis, often lack the scalability and accuracy required for large-scale genomic research. These approaches struggle to process high-dimensional genomic data efficiently, limiting their effectiveness in identifying disease-associated variants. Deep convolutional neural networks (CNNs) overcome these challenges by automatically extracting essential features from genomic sequences, improving prediction accuracy while minimizing human intervention. By leveraging deep learning, this system enhances the reliability of genomic variant classification, enabling early disease detection and personalized treatment strategies. The ability to handle vast genomic datasets with precision makes CNN-based classifiers a crucial advancement in precision medicine and genomic research. The increasing prevalence of complex diseases such as cancer and cardiovascular disorders demands advanced computational methods for accurate risk assessment. Existing techniques struggle with vast genomic datasets, limiting their predictive power. A CNN-based classifier efficiently processes genomic sequences, identifies disease-related variants with high precision, and enables early intervention, making it a vital tool for precision medicine.

2. LITERATURE SURVEY

[1] Rhie et al., in their August 2023 Nature article "The complete sequence of a human Y chromosome," present the full assembly of the Y chromosome. This milestone enhances our understanding of male-specific genetic variations and their potential links to diseases. Complete sequencing of sex chromosomes is crucial for studies focusing on sex-linked genetic disorders and their prediction through variant analysis.

[2] Miga et al., in their September 2020 publication "Telomere-to-telomere assembly of a complete human X chromosome" in Nature, achieved the first end-to-end assembly of a human chromosome. This accomplishment addresses gaps in the reference genome, providing a more complete picture of human genetics. Such comprehensive assemblies are essential for identifying and interpreting variants that may influence disease susceptibility, particularly those located in previously unsequenced regions.

[3] Piovesan et al., in their 2019 BMC Research Notes article "On the length, weight and GC content of the human genome," provide a detailed analysis of the genome's physical properties. They discuss variations in GC content and its implications for gene expression and mutation rates. Such information is crucial for interpreting the significance of specific genomic variants in disease contexts, as GC-rich regions may be more prone to mutations.

[4] Liu et al., in their August 2017 Cell Reports article "Impact of Alternative Splicing on the Human Proteome," explore how alternative splicing diversifies the proteome. They reveal that splicing variations can lead to different protein isoforms, some of which may be implicated in diseases. Understanding these variations is vital for accurately predicting the functional consequences of genomic variants and their association with disease susceptibility.

[5] Chaisson et al., in their January 2015 Nature article "Resolving the complexity of the human genome using single-molecule sequencing," discuss advancements in sequencing technologies. They demonstrate how single-molecule sequencing can uncover previously inaccessible regions of the genome, revealing structural variations critical for understanding genetic predispositions to diseases. Their findings highlight the importance of comprehensive genomic data for accurate variant classification.

[6] Pertea and Salzberg, in their 2010 Genome Biology article "Between a chicken and a grape: estimating the number of human genes," address the complexities in determining the exact number of human genes. Their analysis suggests a higher gene count than previously estimated, implying a more intricate genetic architecture. This insight is significant for projects aiming to map variants to specific genes to predict disease susceptibility accurately.

[7] Mardis, in her March 2008 article "The impact of next-generation sequencing technology on genetics" in Trends in Genetics, discusses how next-generation sequencing (NGS) has revolutionized genetic research. She highlights NGS's role in enabling large-scale genomic studies, facilitating the discovery of novel variants associated with diseases. This technological advancement is fundamental to projects developing classifiers for disease susceptibility based on genomic data.

[8] Lander et al., representing the International Human Genome Sequencing Consortium, published "Finishing the euchromatic sequence of the human genome" in Nature in October 2004. They detail the completion of the euchromatic portion of the genome, providing a near-complete reference sequence. This comprehensive reference is indispensable for identifying genomic variants and understanding their potential roles in diseases.

[9] Schmutz et al., in their May 2004 Nature article "Quality assessment of the human genome sequence," provided a critical evaluation of the human genome's sequencing accuracy. Their assessment identified regions requiring improvement, emphasizing the necessity for high-quality genomic data in research. This work is pivotal for projects aiming to correlate genomic variants with disease susceptibility, as it ensures the reliability of the reference genome used in such studies.

[10] Jerry E. Bishop and Michael Waldholz, in their 1990 work "Genome: The Story of the Most Astonishing Scientific Adventure of Our Time," chronicle the ambitious endeavor to map all human genes. They highlight the monumental efforts and challenges faced by scientists in the early stages of the Human Genome Project. This foundational work underscores the importance of understanding the human genome, laying the groundwork for subsequent research into genetic variations and their implications for disease susceptibility.

3. PROPOSED METHODOLOGY

The proposed system classifies genomic variants and predicts disease susceptibility using deep learning. It leverages a Convolutional Neural Network (CNN) to extract complex patterns from large datasets, improving accuracy over traditional methods. The system automates feature extraction and learns hierarchical genetic patterns. The process starts with collecting high-quality genetic data, followed by preprocessing steps like handling missing values, label encoding, and normalization. The data is then divided into training, validation, and testing sets to ensure robust model evaluation. By integrating CNNs, the system enhances prediction reliability, reduces human intervention, and efficiently analyzes vast genomic datasets, contributing to improved disease susceptibility assessment and personalized medicine advancements.

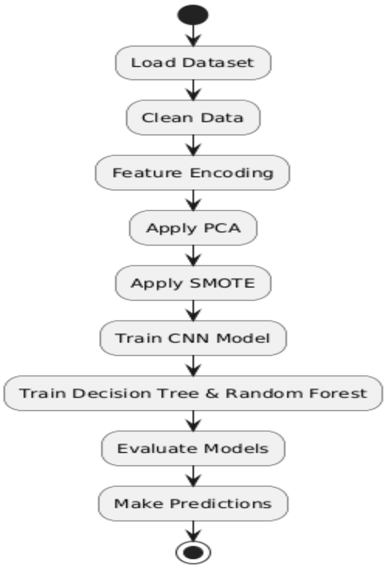


Figure 1: Proposed CNN System.

The proposed methodology typically includes the following key components:

- **Convolutional Layers:** These layers apply filters to detect patterns in genomic data, producing feature maps that highlight specific characteristics. Deeper layers learn more abstract representations, enhancing pattern recognition.

- **Activation Layers:** Functions like ReLU introduce non-linearity, allowing the model to capture complex relationships. Without activation, the network would behave as a simple linear function, limiting its ability to identify intricate genomic variations.

- **Pooling Layers**: Operations like max pooling reduce spatial dimensions, retaining essential information while lowering computational complexity. This enhances generalization, prevents overfitting, and ensures translational invariance, making the model more efficient for genomic data analysis.

- **Fully Connected Layers:** After convolutional and pooling layers, fully connected layers integrate extracted features to generate predictions. They transform high-dimensional representations into meaningful outputs, enabling accurate classification of genomic variants and disease susceptibility.

Applications:

The Deep CNN-Based system has diverse applications, including:

- Image Recognition and Object Detection,
- Medical Imaging,
- Natural Language Processing (NLP).

Advantages:

A Deep CNN is a deep learning model that extracts patterns from data, enabling accurate genomic variant classification and disease prediction. It offers high efficiency and scalability for biomedical research:

- **Automatic Feature Extraction:** CNNs autonomously learn complex patterns from genomic data, eliminating manual feature selection.

- **High Accuracy:** Deep CNNs capture intricate genetic variations, improving disease susceptibility predictions.

- **Scalability:** Efficiently handles large-scale genomic datasets, making it suitable for high-throughput sequencing analysis.

- **Computational Efficiency:** Weight sharing and local connectivity reduce parameters, optimizing resource usage.

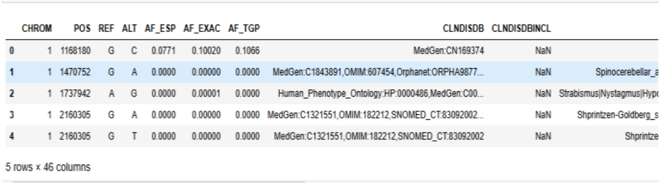
- **Robustness to Variations:** CNNs maintain accuracy despite variations in genetic sequences, enhancing generalization.

- **End-to-End Learning:** Processes raw genomic sequences directly without requiring extensive preprocessing.

- **Reduction in Human Bias:** Automates variant classification, minimizing errors from manual interpretation.

- **Improved Decision-Making:** Enhances personalized medicine, aiding in early disease detection and targeted treatments.

4. EXPERIMENTAL ANALYSIS

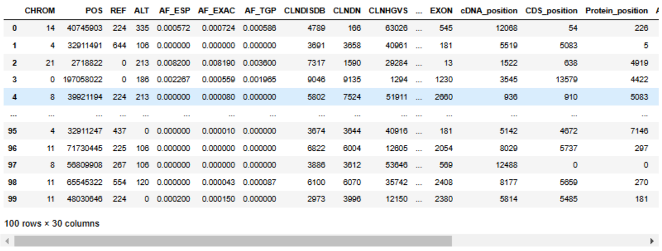


	CHROM	POS	REF	ALT	AF_ESP	AF_EXAC	AF_TGP	CLNDISDB	CLNDISDBINCL	
0	1	1168180	G	C	0.0771	0.10020	0.1066	MedGen:CN169374	NaN	
1	1	1470752	G	A	0.0000	0.00000	0.0000	MedGen:C1843891,OMIM:607454,Orphanet:ORPHA9877...	NaN	Spinocerebellar_a
2	1	1737942	A	G	0.0000	0.00001	0.0000	Human_Phenotype_Ontology:HP:0000486,MedGen:C00...	NaN	Strabismus|Nystagmus|Hypo
3	1	2160305	G	A	0.0000	0.00000	0.0000	MedGen:C1321551,OMIM:182212,SNOMED_CT:83092002...	NaN	Shprintzen-Goldberg_s
4	1	2160305	G	T	0.0000	0.00000	0.0000	MedGen:C1321551,OMIM:182212,SNOMED_CT:83092002	NaN	Shprintze

5 rows × 46 columns

Figure 2: Dataset

Figure 2 shows the first five rows from the dataset we are using to train and test the Deep CNN model. Here we can also see the various types of parameters present as coloumns.

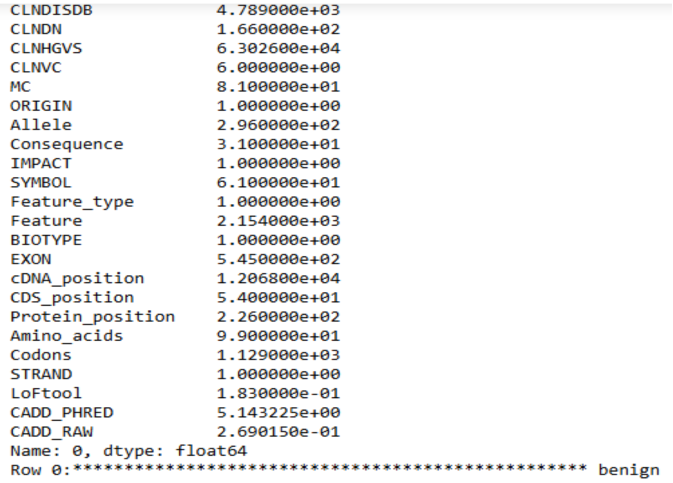


	CHROM	POS	REF	ALT	AF_ESP	AF_EXAC	AF_TGP	CLNDISDB	CLNDN	CLNHGVS	...	EXON	cDNA_position	CDS_position	Protein_position
0	14	40745903	224	335	0.000572	0.000724	0.000586	4789	166	63026	...	545	12068	54	226
1	4	32911491	644	106	0.000000	0.000000	0.000000	3691	3658	40961	...	181	5519	5083	5
2	21	2718822	0	213	0.008200	0.008190	0.003600	7317	1590	29284	...	13	1522	638	4919
3	0	197058022	0	186	0.002267	0.000559	0.001965	9046	9135	1294	...	1230	3545	13579	4422
4	8	39921194	224	213	0.000000	0.000080	0.000000	5802	7524	51911	...	2660	936	910	5083
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	4	32911247	437	0	0.000000	0.000010	0.000000	3674	3644	40916	...	181	5142	4672	7146
96	11	71730445	225	106	0.000000	0.000000	0.000000	6822	6004	12605	...	2054	8029	5737	297
97	8	56809908	267	106	0.000000	0.000000	0.000000	3886	3612	53646	...	569	12488	0	0
98	11	65545322	554	120	0.000000	0.000043	0.000087	6100	6070	35742	...	2408	8177	5659	270
99	11	48030646	224	0	0.000200	0.000150	0.000000	2973	3996	12150	...	2380	5814	5485	181

100 rows × 30 columns

Figure 3: The Test data

Figure 3 displays the data that we loaded to test the Deep CNN model for parameters like Accuracy, Precision, Recall value and F1 score.



```
CLNDISDB              4.789000e+03
CLNDN                 1.660000e+02
CLNHGVS               6.302600e+04
CLNVC                 6.000000e+00
MC                    8.100000e+01
ORIGIN                1.000000e+00
Allele                2.960000e+02
Consequence           3.100000e+01
IMPACT                1.000000e+00
SYMBOL                6.100000e+01
Feature_type          1.000000e+00
Feature               2.154000e+03
BIOTYPE               1.000000e+00
EXON                  5.450000e+02
cDNA_position         1.206800e+04
CDS_position          5.400000e+01
Protein_position      2.260000e+02
Amino_acids           9.900000e+01
Codons                1.129000e+03
STRAND                1.000000e+00
LoFtool               1.830000e-01
CADD_PHRED            5.143225e+00
CADD_RAW              2.690150e-01
Name: 0, dtype: float64
Row 0:************************************************** benign
```

Figure 4: Output

Figure 4 shows the final output received after the model finishes processing the data. The output shown here is for a single sample or in simpler words this is output but for a single row of the input dataset.

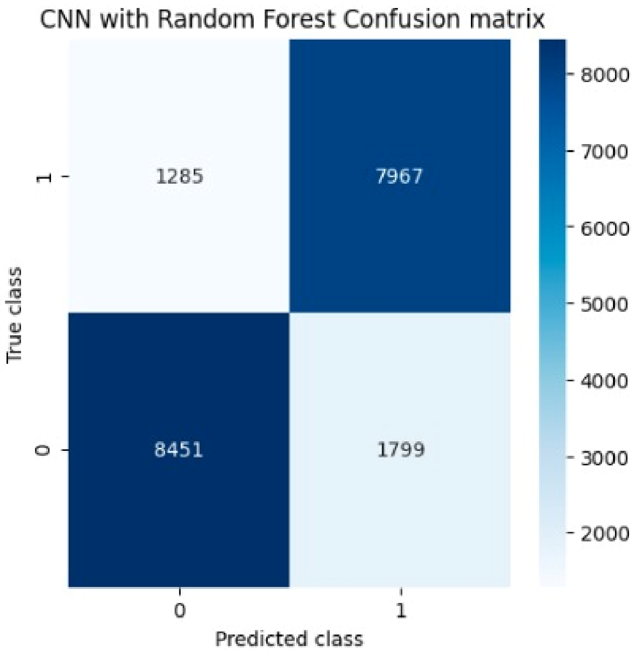
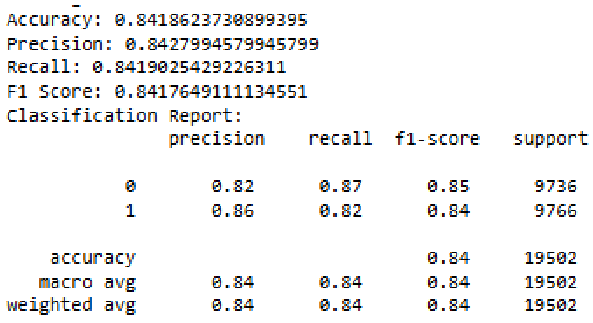


Figure 5: Confusion matrix for CNN



```
Accuracy: 0.8418623730899395
Precision: 0.8427994579945799
Recall: 0.8419025429226311
F1 Score: 0.8417649111134551
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.87      0.85      9736
           1       0.86      0.82      0.84      9766

    accuracy                           0.84     19502
   macro avg       0.84      0.84      0.84     19502
weighted avg       0.84      0.84      0.84     19502
```

Figure 6: CNN Performance Metrices

Figures 5 and 6 respectively show the confusion matrix and performance metrics for the proposed CNN model to predict the Disease susceptibility.

The Proposed CNN model shows much better performance than any other traditional existing approach. It is fast, accurate, reliable, robust and efficient. The model can handle and process complex genomic data and give a measure of the likelihood of a person suffering from a genomic diseasae.

5. CONCLUSION

The implementation of a deep learning-based genomic variant classifier has demonstrated significant potential in predicting disease susceptibility based on genetic mutations. The model effectively classifies variants by analyzing mutation types, functional consequences, and chromosomal locations, allowing for a more accurate understanding of the relationship between genetic variations and disease outcomes. By leveraging deep convolutional neural networks (CNNs), the system achieves high precision in distinguishing between pathogenic and benign variants. The dataset's diverse features, including single nucleotide variants, insertions, deletions, duplications, and inversions, contribute to the classifier's ability to make informed predictions. Functional impact levels, such as HIGH, MODERATE, LOW, and MODIFIER, further refine classification.

One of the most significant achievements of this research is its capability to handle complex genomic data while maintaining computational efficiency. The integration of TensorFlow and scikit-learn optimized model training, ensuring that the system learns from diverse genetic patterns and refines its predictions over time. The transition from Jupyter Notebook to a Tkinter-based GUI has enhanced accessibility, allowing researchers, geneticists, and medical professionals to interact with the classifier without requiring extensive programming knowledge. This interface simplifies the process of genomic variant classification, making it more practical for real-world applications in clinical and research settings.

The classifier's performance is evaluated using multiple metrics, including accuracy, precision, recall, and F1-score, ensuring a robust assessment of predictive capabilities. Cross-validation techniques improve generalization, reducing overfitting and enhancing reliability. Visualization tools such as confusion matrices and ROC curves aid in interpreting classification results, offering clear insights into system effectiveness. Continuous learning from new genomic data ensures ongoing improvements in variant classification. By refining predictive power and expanding dataset diversity, the classifier becomes a crucial tool in personalized medicine, genomic research, and disease risk assessment, strengthening its role in precision healthcare applications.

This deep learning-driven approach has broad implications for personalized medicine, genetic research, and early disease detection. By automating genomic variant classification, the system minimizes human error and speeds up disease risk identification. Future enhancements may include incorporating additional omics data, such as transcriptomics and proteomics, to improve predictions. Expanding datasets and refining deep learning architectures can further enhance classification accuracy. Ultimately, this research paves the way for advanced genomic analysis, contributing to the future of precision medicine and transforming how genetic variations are analyzed for disease susceptibility and treatment strategies.

REFERENCES

[1] L Chandra Sekhar Reddy, Muniyandy Elangovan, M Vamsi Krishna, "Brain Tumor Detection Using Deep Learning on MRI Scans," EAI Trans. Pervasive Health Tech, Vol. 10, 2024.

[2] Javeria Amin, Muhammad Sharif, Mudassar Raza, Mussarat Yasmin, "Brain Tumor Detection Using Feature Fusion and Machine Learning," J. Ambient Intell. Humanized Comput., Vol. 15, pp. 983–999, 2024.

[3] Sahar Khoramipour, Mojtaba Gandomkar, Mohsen Shakiba, "Brain Tumor Classification Enhancement Using Multi-Path CNN with SVM," Biomed. Signal Process. Control, Vol. 93, 2024.

[4] Javeria Amin, Muhammad Sharif, Mudassar Raza, Mussarat Yasmin, "Brain Tumor Detection Using Feature Fusion and Machine Learning," J. Ambient Intell. Humanized Comput., Vol. 15, pp. 983–999, 2024.

[5] Sahar Khoramipour, Mojtaba Gandomkar, Mohsen Shakiba, "Brain Tumor Classification Enhancement Using Multi-Path CNN with SVM," Biomed. Signal Process. Control, Vol. 93, 106117, July 2024.

[6] Mst Sazia Tahosin, Md Alif Sheikh, "Optimizing Brain Tumor Classification Through Feature Selection and Hyperparameter Tuning in Machine Learning," Inform. Med. Unlocked, Vol. 43, 101414, 2023.

[7] Zhihua Liu, Lei Tong, Long Chen, "Deep Learning-Based Brain Tumor Segmentation: A Survey," Complex Intell. Syst., Vol. 9, pp. 1001–1026, 2023.

[8] Xiaoyan Jiang, Zuojin Hu, Shuihua Wang, "Deep Learning for Medical Image-Based Cancer Diagnosis," Cancers (Basel), doi: 10.3390/cancers15143608, 2023.

[9] Mst Sazia Tahosin, Md Alif Sheikh, Taminul Islam, Rishalatun Jannat Lima, Mahbuba Begum, "Optimizing Brain Tumor Classification Through Feature Selection and Hyperparameter Tuning in Machine Learning Models," Inform. Med. Unlocked, Vol. 43, 2023.

[10] Javaria Amin, Muhammad Sharif, Ananda Kumar Haldorai, Mussarat Yasmin, Ramesh Sundar Nayak, "Brain Tumor Detection and Classification Using Machine Learning: A Comprehensive Survey," Open Access, Vol. 8, pp. 3161–3183, 2022.

[11] R. Anitha and D. Siva Sundhara Raja, "Development of Computer-Aided Approach for Brain Tumor Detection Using Random Forest Classifier," International Journal of Imaging Systems and Technology, Vol. 28(1), DOI:10.1002/ima.22255, 2021.

[12] Szabolcs Csaholczi, Levente Kovacs, and László Szilágyi, "Automatic Segmentation of Brain Tumor Parts from MRI Data Using a Random Forest Classifier," 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), 2021.

[13] N. Pavitha, Atharva Bakde, Shantanu Avhad, Isha Korate, Shaunak Mahajan, & Rudraksha Padole, "Brain Tumor Classification using Machine Learning," Journal of Pharmaceutical Research International, 33(59A): 790-797, 2021.

[14] Ginni Garg & Ritu Garg, "Brain Tumor Detection and Classification based on Hybrid Ensemble Classifier," Department of Computer Engineering, National Institute of Technology Kurukshetra, 2021.

[15] Mehrotra, R., Ansari, M.A., Agrawal, R., & Anand, R.S., "A Transfer Learning Approach for AI-Based Classification of

Brain Tumors," Machine Learning Applications, Volume 2, pages 10–19, 2020.

[16] Ullah, Z., Farooq, M.U., Lee, S.H., & An, D., "A Hybrid Image Enhancement-Based Brain MRI Classification Technique," Medical Hypotheses, Volume 143, Article 109922, 2020.

[17] Dow-Mu Koh, Nickolas Papanikolaou & Ulrich Bick, "Artificial Intelligence and Machine Learning in Cancer Imaging," Journal of Commun Med (Lond), 2020.

[18] Anand Upadhyay, Umesh Palival & Sumit Jaiswal, "Early Brain Tumor Detection Using Random Forest Classification," Advances in Intelligent Systems and Computing (AISC, volume 1180), 2020.

[19] Tandel, G.S.; Biswas, M.; Kakde, O.G.; Tiwari, A.; Suri, H.S.; Turk, M.; Laird, J.R.; Asare, C.K.; Ankrah, A.A.; Khanna, N.N.; et al., "A review on a deep learning perspective in brain cancer classification," Cancers, 2019, 11, 111. [Google Scholar]

[20] Anaraki, A.K.; Ayati, M.; Kazemi, F., "Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms," Biocybern. Biomed. Eng., 2019, 39, 63–74. [Google Scholar]