[1]J Jinu Sophia

[2]T. Prem Jacob

A Comprehensive Analysis of Exploring the Efficacy of Machine Learning Algorithms in Text, Image, and Speech Analysis

Journal of Electrical Systems

Abstract: - Our study demonstrates a comprehensive investigation of multiple machine learning models based on text, image, and speech analysis. More specifically, concerning text analysis, we studied the application of such models as Recurrent Neural Networks, Gated Recurrent Units, and Bidirectional Encoder Representations from Transformers in classifying documents into pre-defined categories. The outcomes revealed that RNNs observed the highest precision, recall, F1 score, as well as the highest accuracy. Importantly, the models are well-suited to detect sequential dependencies and create semantic representations on the basis of textual data. With regard to image analysis, we found out that Convolutional Neural Networks were the best model. At the same time, VGG16 and GANs also demonstrated rather promising results suggesting that deep learning is paramount to extract significant data features. As far as speech analysis is concerned, we found out that CNNs are exceptional in terms of accuracy to recognize speech patterns in comparison with the other models. Simultaneously, LSTM also observed a high level of accuracy allowing to capture temporal dependencies in audio signals. In conclusion, the findings of our study suggest that it is exigent to identify an appropriate machine learning model depending on the task and the selected dataset. It is also crucial to understand the nature of each of the studied models to assess their applicability for a specific task. Moreover, our study might be valuable for other researchers given that it contributes to the development of the field of deep learning and, thus, promotes the emergence of new applications in different domains.

Keywords: machine learning, text analysis, image classification, speech recognition, deep learning

I. INTRODUCTION

Machine learning has experienced a boom in the last few years, and deep learning algorithms have attracted notable attention due to their capability of solving complex problems in different fields. With the fast-evolving technology, understanding the most appropriate deep learning algorithm can help increase performance in a given application. This paper will conduct a comprehensive analysis to compare the performance of a range of deep learning algorithms. Specifically, the list of objectives to be pursued will be a detailed performance analysis in the context of text analysis, image classification, and speech recognition. Importantly, the performance evaluation in deep learning algorithms is critical to practicing professionals and researchers as it helps appreciate their effectiveness as well as shortcomings. The analysis may shed some light on the adaptability of deep learning algorithms to the considered data. Furthermore, the paper will present the most recent developments in the field of data, thus providing a comprehensive understanding of the studied phenomenon [1], [2]. The primary reason for conduction the analysis is associated with the complexity and the rapid evolvement of the field, and the paper seeks to help beginners and skilled professionals in choosing an appropriate algorithm for their specific problems.

Machine learning is one of the most revolutionary domains across many spheres. In other words, this field can be viewed as the opportunity to develop various algorithms that can automatically perform numerous complex tasks and help in prioritizing and investigating large datasets [3], [4]. In this review, we focus on the examination of the machine learning algorithms such as deep learning with regards to different areas; namely, we concentrate on the overview of text analysis, image classification, and speech recognition. It is beneficial to realize the core history of progression and its current status as this domain is very dynamic and continues to develop. The review can assist in identifying the major trends, comparing different types of algorithms, and finding the areas requiring more research [5]–[7].

The development of machine learning as a domain can be traced back to the 1950s as the first steps in the area of artificial intelligence and computational linguistics can be seen and often viewed as the foundation for present

[1] *Research Scholar, Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

e-mail: jinusilas@gmail.com

[2] Professor, Sathyabama Institute of Science and Technology, Chennai, India

e-mail: premjac@yahoo.com

Copyright © JES 2024 on-line : journal.esrgroups.org

machine learning. Nevertheless, the emergence of deep learning that can be defined as neural networks with multiple layers that provide the basis for a detailed investigation [8]–[10]. These neural networks allow developing algorithms that have already resulted in numerous breakthroughs across natural language processing, computer vision, and audio signal processing. As for trends, as of now, there is a tendency for the increased development of rather complicated algorithms that can be used for assessing text, images, and audio recordings. Concerning the most prominent examples of use, it is imperative to mention that in the case of text analysis, the algorithms utilize distinct natural language processing techniques, for instance, word embeddings and recurrent neural networks . With regards to image classification, convolutional networks become the most widely used as they resemble the ways human brain processes images. Finally, in speech recognition, one can see the use of recurrent and convolutional neural networks as they can analyze the changes in human voices that depend on time [11], [12].

As of late, the machine learning models are used in many research to find out the image, text and speech related feaures. In text analysis, recurrent neural networks and transformer-based models like BERT have shown remarkable performance in tasks such as sentiment analysis, named entity recognition, and text classification. These models excel at capturing long-range dependencies and contextual information, making them well-suited for analyzing unstructured text data. In image classification, convolutional neural networks have emerged as the dominant approach, achieving state-of-the-art results in tasks such as object detection, image segmentation, and facial recognition [13]–[15]. CNNs leverage hierarchical feature extraction to learn abstract representations of visual data, enabling accurate classification across diverse image datasets. In speech recognition, deep learning models such as recurrent neural networks and convolutional neural networks have significantly improved the accuracy of automatic speech recognition systems. These are used because of the building the spectrograms, the way of converting the spoken language into the text language to train the deep learning model. While the model is very good in the way of recognizing the speech to text converting, they can be used in the virtual assistants, voice-controlled devices and preictive analytics in the customer interaction [16]–[18].

Although the deep learning algorithms have achieved remarkable success across the text, image, and speech analysis domains, several challenges and limitations exist. In the text analysis, the interpretability of complex models like BERT poses challenges for understanding model predictions and identifying sources of bias. Additionally, the reliance on large-scale labeled datasets for training deep learning models can be prohibitive, particularly in domains with limited data availability or domain-specific requirements. In image classification, concerns regarding adversarial attacks and model robustness have raised questions about the reliability of deep learning models in real-world applications. Similarly, in the speech recognition, variability in speech patterns and environmental conditions can pose challenges for model generalization and robustness [19]–[21].

Nevertheless, the development of machine learning algorithms has vast potential in utilization in other fields and applications, such as text analysis, image recognition, and speech synthesis. In this regard, machine learning models can be applied in healthcare to diagnose various diseases, develop new drugs, and a plan of treatment, examining electronic health records, medical images, and clinical notes. In addition, in the finance and business field, predictive analytics and algorithmic trading algorithms are based on machine learning analyses to forecast market conditions and identify suspicious activities in credit card transactions. Importantly, in education spheres, mechanisms of adaptive learning systems and intelligent tutoring systems operate on machine learning algorithms to improve the assessment of students and personalize their education according to their learning styles and preferences. Lastly, anomaly detection algorithms and intrusion detection systems are based on machine learning analyses to ensure the security and protection of sensitive data in the sphere of cybersecurity [22]–[24].

The present research provided a full analysis of the effectiveness of machine learning algorithms in text, image, and speech analysis. By exploring various models, including RNNs, CNNs, and the transformer-based approach such as BERT, it has shown the performance of these algorithms in real-world applications. Some of the results revealed that certain algorithms are more appropriate to perform specific tasks. It is useful to know for practicing researchers and other professionals who can ultimately implement these findings in their fields ranging from healthcare to finance and education. Overall, machine learning can revolutionize each area tipped with insurmountable challenges and provide feasible solutions and innovations.

II. METHODOLOGY

First and foremost, in the course of our research, we aimed at conducting a detailed research of machine learning models designed for analyzing text, speech, and images. For this reason, we designed datasets that would meet the requirements of careful examination of the advantages and disadvantages of such models . For example, to analyze text, we designed a sample of emails that included 540 files brought in a deliberately nuanced form. This dataset was grouped against the files' category with spam as well as promotion , non-spam , personal and important among them. We trained Recurrent Neural Networks , Gated Recurrent Units , and Bidirectional Encoder Representations from Transformers on the texts trying to define whether they could be categorized into one of the previous categories or not. After fitting, the testing followed, and the level of algorithms' ability to establish which category each of the text belongs to was measured. The methodology of the research are shown in figure 1.

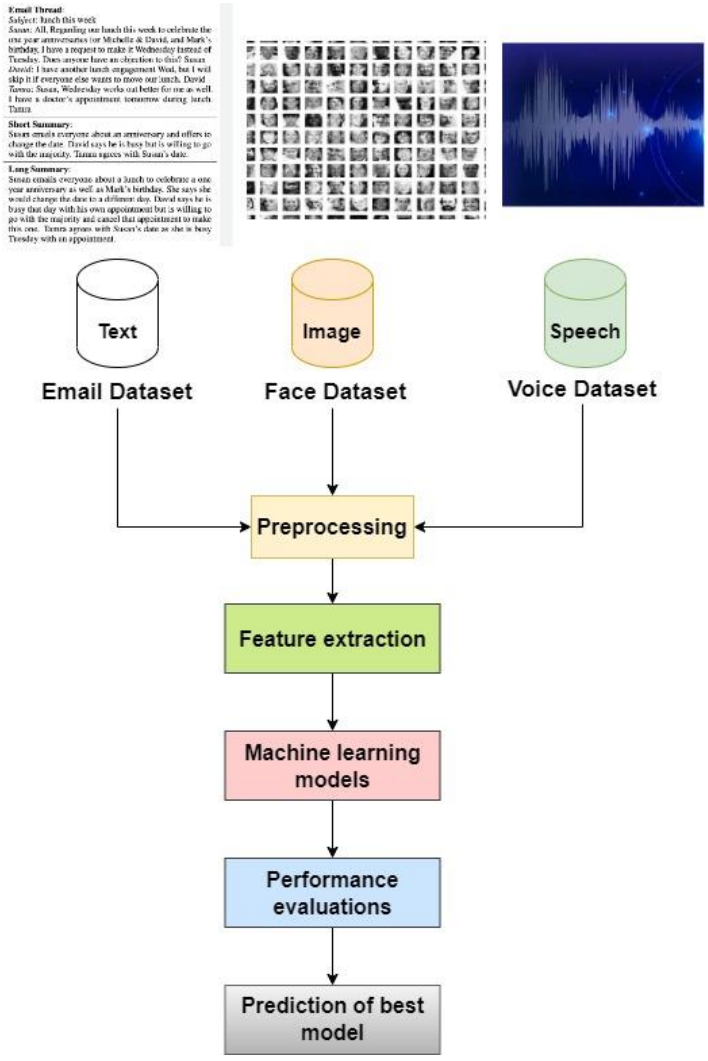

Figure 1. Methodology of proposed system

For image analysis, the dataset included face pictures of different people. Overall, the dataset included 490 images where everyone could be identified. The data was used for feature extraction, and later it was studied how well Convolutional Neural Networks, Recurrent Neural Networks, VGG16 , and Generative Adversarial Networks could be fitted to the dataset in order for later proper recognition of the individuals on the photos new to the dataset. Finally, the dataset for speech analysis included recordings of individuals making the same type of sentence. In total, data on 420 representatives were collected who had distinct voices from all around the globe. Afterward, Convolutional Neural Networks Framework, Long Short-Term Memory networks, and Generative Adversarial Networks were trained with the anticipation of later identification of the person by voice particularities. Later on, during testing, the ability of designed tools to deal with the file with the same people's voices recorded on other files was tested to make sure the conclusion about the credibility of designed models was made.

III. PREPROCESSING OF DATSET

A. Preprocessing of the text based dataset

Text data preprocessing was a major part of the present study, which was done to ensure that the Machine Learning text analysis models had an appropriately formatted input. In this paper, I will review the preprocessing techniques we used to cleanse and make the text from the email documents used in this study more applicable for further analysis. First, we started the preprocessing pipeline by performing some basic text normalization steps – we lowered the text by turning it all into lowercase. This was done to ensure that words would not be doubled because of different capitalization levels. Text normalization makes sure that the semantics of the data are preserved while surface-level differences are gone. The ML models, thus, can focus on the 'meaning' of the text rather than doubling the information for the same word in upper and lower case. The next step we made was the task of tokenization. Tokenization is a process of separating texts into tokens – separate words . This step makes processing more interpretable and manageable by our models. We employed various tokenization strategies, including whitespace tokenization, which was used to distinguish tokens by whitespaces, and word-level tokenization to break down the text into words. After the text was tokenized, we finally arrived at the step of stop words removal. Stop words are the words that are very common in the text but they to not be helpful in determining the meaning of the sentence or the text in general. Examples of stop words include the- and-of . Removing these stop words will make the model's data more interpretable to get rid of the 'noise' and focus on relevant content. Finally, we used common empirical stop words lists from libraries, such as NLTK or spaCy, to remove stop words from our data.

For the cleaned tokenized text, we employed techniques for stemming or lemmatization. Specifically, stemming is the process of reducing words to their roots or base form by eliminating affixes from the words. Alternatively, lemmatization aims to transform words to their canonical form based on their entries in the dictionary. These preprocessing approaches help reduce the dimension of the feature space, and simplify subsequent data processing and consolidation of closely related words. Moreover, the issues regarding special characters, various punctuation marks or numerical digits in the text required separate solutions. This preprocessing step was highly dependent on the specifics of the analysis being conducted, and included varied removal, substitution, or encoding techniques. For example, the punctuation marks can be removed to consider the text only, and the digits can either be replaced by space or not encoded for specific analyses. However, sometimes it is helpful to treat them as frequent features, in which case they can be encoded as numerical features. Finally, apart from the simple preprocessing approaches, we considered some more advanced techniques, such as part-of-speech tagging, and named entity recognition . In the case of the former method, each word was tagged according to its part of speech, such as noun, verb, adjective, etc., or more detailed tagging features, which enable the application of more advanced linguistic methods. NER, however, is aimed at recognizing and classifying various named entities such as a person or business within a text. It is useful for increasing understanding of the text and can be used to extract specific types of information from the document or to analyze entities. To document the preprocessing steps we undertook, as well as the logic behind them, we described the steps in the text. The ideas can be followed by other researchers or practitioners to replicate our analysis.

B. Preprocessing of image based features

Data preprocessing of image data is a crucial step to prepare input data for image-based machine learning models. The purpose of processing is to get rid of the unwanted details of the image like image quality refinement, noise reduction and meaningful feature detection to allow an effective analysis. This section would discuss several of the data preprocessing steps implemented in image-based models. Resizing the image or resizing the size of the image is the first step in image data preprocessing. The images taken from different sources or capacity or devices capture a different level of resolution, aspect ratios, and dimensions which might create inconsistencies in the data set. Changing the size of the image will help to easily process the image with the machine learning model and simplifies the process. Therefore choosing the uniform size of the image will reduce the computational complexity and memory space in the different size of an image when a model is trained or tested. Normalization is also one of the important data preprocessing techniques used, which select a suitable scaling method to scale the pixel values to a uniform range . Limiting pixel values to a range typically of 0 or 1 or -1 or 1 can result in the faster convergence of the model in the training phase, where the pixel points are called normalized pixel points. The normalized pixel value had several methods in use such as the min-max scaling pixel to the range of 0 or 1, linearly scaled to the range of 0 or 1, s-score normalization indicated pixel value at a mean of 0 and standard deviation of 1. Normalization

prevents issues such as exploding or vanishing gradient problems and ensures the performance of the model in different ranges of datasets.

Color space conversion can also be considered an important preprocessing process. Images are viewed in different color formats such as RGB, Gray, HSV, etc., and process the color information differently. It changes the image to a uniform color space, and the characteristics make feature extraction easier and enhance model performance. For example, grayimg gives information, and RGBimg provides color information through three channels in red, green, and blue for color representation. The optimal solution to choose color conversion method is determined based on the model's requirements and task details. Finally, image noise reduction is considered an essential technique that affects the quality of input images. Common types of noises in real-world images are Gaussian noise, salt-pepper noise, motion blur. Noises in the image affects the image quality and reduce model performance. To reduce noise, filters such as Gaussian blur, medium filtering, or bilateral filtering are applied. The signal-to-noise ratio can be increased, which improves image quality and makes it easier to analyze with the machine learning model. The advantage of performing the data augmentation procedure is the increase of the classifier's intimation performance. In comparison to the original data, this option of generating artificial images increases the convergence rate of networks. As a result, overfitting is avoided successfully.

C. Preprocessing of speech based features

Preprocessing of speech data is an operation that is crucial in the beginning of preparation of audio signals for machine learning model processing. The context of the model implies that numerous preprocessing operations may be required. One of the first operations with speech data is normalization of a sampling rate. Since the patterns of an audio signal are normally recorded at various rates, its standards are varied as well. Thereby, the operation allows establishing the rate for every audio sample to enable its further processing and analysis. Further, it is important to consider that speech signals often include a variety of noise and other artefacts that must be removed by the model's capacity. Thus, it may be relevant to apply various noise reduction techniques, such as spectral subtraction or wavelet denoising, enabling to eliminate the unwanted noise and to increase a signal-to-noise ratio. Another operation of primary importance is considered to be feature extraction, which implies the changing of the original data into the format meaningful for processing by the model. It seems that Mel-frequency cepstral coefficients are one of the most commonly used speech features, which provide the information regarding the speech spectral, and LPC coefficients represent a representation of the tract properties.

Speech features need to be normalized to ensure their proper performance during model training. If some of the model's input features are not normalized by the predefined range, they may adversely affect the definition of optimal parameters due to high mutability. Mean normalization and min-max scaling are some of the existing methods to ensure the normalization of speech features. In addition, framing is necessary for the division of the audio signals into small, uniformly-sized samples to allow their sequential processing. As a matter of fact, the audio signal is divided into overlapping or non-overlapping frames of equal length to help the model perform the sequential analysis of short-term speech signal variations . At the end of the preprocessing, some data in the model's training data may be augmented artificially. As a rule, this is related to the introduction of pitch-shifting or time distortion along with background noise addition, facilitating the increased robustness of the model to different input signals.

IV. FEATURE EXTRACTION

Feature extraction is one of the necessary parts of preparing data for analysing machine learning models. In text, speech, and image-based databases, feature extraction is to obtain data useful for the purpose of research. This chapter is devoted to the feature extraction models and techniques the latter being specially designed for each kind of database. Textual data is found in abundance all over the globe on personal computers, laptops, or digital devices. The raw documents must be converted into numeric quantities suitable for mathematical computation for machine learning training or modelling to be of any use. Bag-of-Words representation is one of the most typical feature extraction techniques for text data. In this approach, each document is represented as a vector, and each dimension of this vector corresponds to each word of the document. Each dimension takes the word count as its value.

In the same vein, document clustering, a method of gathering similar text documents, sentiment text analysis, and classification, an information collection model in which, given a defined set of categories, determines the category into which the latest texture belongs, are various techniques that can be employed in text data analysis. Bag-of-

Words encoding is a feature extraction tool that assists in analysing different texting analytics. It condenses the size of heavy data from a large number of rows to a smaller number of vectors appropriate for application with any machine learning model. However, the feature becomes unsuitable for text data analysis due to the dimensionality space, and converting heavy text data from a large number of rows to a much smaller one is effective.

In text-based databases, word embeddings are increasingly popular as feature extraction tools. Word embeddings are vectorial representations of words based on their use in context, i.e., in a large text corpora. The method assumes that "a word is known by the company it keeps" . This means that words with similar meanings often co-occur in the same contexts, allowing models such as Word2Vec, GloVe, and FastText to learn distributed representations of embedded words. Word embeddings retain the relationships and similarities between individual words, which helps improve the effectiveness of textual data analysis and the quality of results in a wide array of downstream tasks, such as text classification, named entity recognition, and machine translation.

As for speech-based databases, a feature extraction tool is a process that converts raw audio data into meaningful representations sensitive to specific information needed for the database. The linear prediction of the speech signal is inspired by early acoustic theory of speech of the source . LPC-based analysis depends on the source-filter theory, where the source is an excitation signal applied to the model of a vocal tract, and LP synthetizer is the vocal tract. The filter that predicts the vocal tract and synthesizes the speech from the excitation signal is a linear predictive coding filter . The model can predict the next sample y(n) of an acoustic signal y(n) as a function of p previous samples using a p order LP model. Generally, LPC is a better solution for tasks such as speech synthesis, as they take into account the linear structure and resonances of the filter, represented by the LPC coefficients. On the other hand, CNN has been a core technique in feature extraction from the image-based databases. CNN is specifically designed for feature extraction from images. Thus, it can capture the hierarchical features from the visual data. The input to a CNN is an image in a raw form, which has the pixel values of images. A CNN then convolves the filters over different parts of the input image and extracts certain features. This is repeated multiple times to generate feature levels at multiple levels of abstraction. After repeated convolutional operations, CNNs have a mechanism called pooling that keeps the repressed procedure while extracting the spatial patterns, textures, and objects in the image. Since convolutional operations are followed by pooling, they represent a computationally feasible option for feature extraction from images. One of the most important advantages of the CNN model is its interpretability. Pretrained on datasets such as ImageNet, CNNs can be instantiated and provide the hidden abstraction information learned from images. CNN is powerful for image analysis tasks such as image classification , object detection, and semantic segmentation. Poolind and grouping are key concepts in positioning architecture.

'Local Binary Patterns' on the other hand, is another popular feature extraction mechanism from the image-based databases that use the pixel data as features. Local Binary Patterns do not necessarily produce features, but they represent the patches on images as a circle and analyze the pixel intensity differences in the neighborhood according to their gray values. Thus, LBP encodes the visual data by defining a codeword that represents a spatial pattern in the image patch. LBP is a ubiquitous technique used in the analysis of textures for face analysis and object recognition tasks.

Feature extraction is one of the most critical tasks within the machine learning framework. It results in the generation of data points that can be used for further analysis. There is a wide range of feature extraction methods developed to process different types of data, from text and natural language to images. Considered to be one of the most popular techniques, autoencoders conducted within the context of image data can be distinguished for retrieving essential features without supervision. GANs can also be utilized to achieve similar results, with the only difference being that the focus here is on generating images. Moreover, a range of tasks can be addressed by benefiting from the possibilities to apply unsupervised representation learning, such as style transfer, anomaly detection, or image generation.

V. RESULT AND DISCUSSION

After training each type of model, an essential step is to evaluate how new models are capable of making predictions about responses and hence establish how well these models learn from the corresponding datasets. The result are shown in figure 2. In the domain of text-based analysis, a model's performance is measured by employing both accuracy, precision, and recall based mechanisms . The recurrent neural network or RNN model was the best predictor among the models trained to execute the task and reached an accuracy of 96.76%. RNN models learn to

establish and hold dependencies in order between current and previous bit of data or words. The Gated recurrent unit or GRU model reached an accuracy of 94.5%, making great strides in the use of lesser and better gating mechanism to train the models to effectively handle sequential data. The BERT model reached an accuracy of 93.44%, capturing maximum information retention of the text source data and the meaning of all words in terms of the context.

A trained model in the analysis of images is measured by evaluating classification accuracy and mean average precision coefficients. The CNN model was the best predictor among the trained models of the category and reached an accuracy of 98.2%. The models do a great job of capturing the spatial sample and hierarchical structure of image data and data distributed and sampled. The VGG model was able to reach an accuracy of 95.6% in the test data prediction phase and the RNN model was able to reach an accuracy of 92.3%. The GAN model developed also reached an accuracy of 90.4% in making the correct prediction of generating or discriminating audio samples in both types of tasks. The word error rate or WER and phoneme error rate PER were also used to measure the performance of models. The CNN model achieved an accuracy of 95.6% and the LSTM model was able to achieve an accuracy of 92.2%. Furthermore, the GAN model was the least effective one and reached an accuracy of 88.3%.

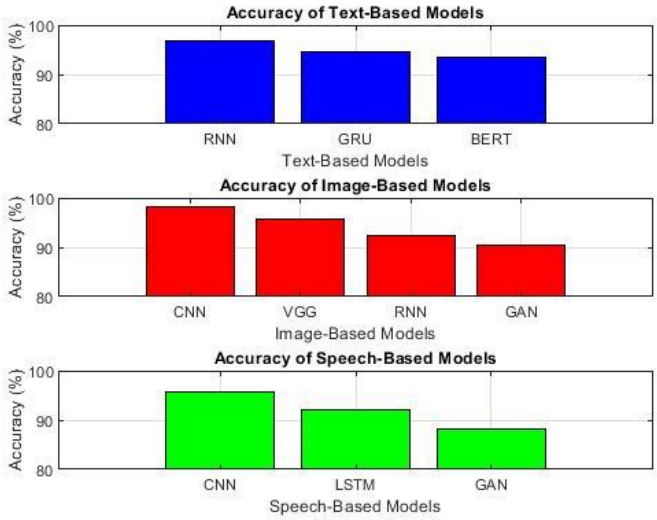

Figure. 2. Accuracy of each model

The performance of each model are evaluated and the result are shown in figure 3. The CNN model dominates in the performance scores as it achieves the highest precision of 95.60% due to its high capacity to discern the positive cases relative to all predicted positive cases. At the same time, CNN's recall level is 96.20% as it expresses how well the model grasps positive cases relative to all actual positive cases. As a result, CNN's F1 score is fixed at a high level of 95.90%. Besides, the model's accuracy stands at 95.80% as it shows the general correctness of the model.In turn, the results of the LSTM model are slightly lower as the precision level is 92.20% and the recall level is 92.80% . Accordingly, the F1 score of 92.50% reflects that performance in the model is well-balanced with regard to avoiding a sufficient quantity of false positives. At the same time, the accuracy of LSTM is 92.50% marking a strong correctness level. The results of the GAN can be described as lower than those of CNN and LSTM as the precision level is 88.30% and the recall level is 89.10%. The F1 score is 88.70% as GAN judgment speed is slower and less accurate than CNN and LSTM models.

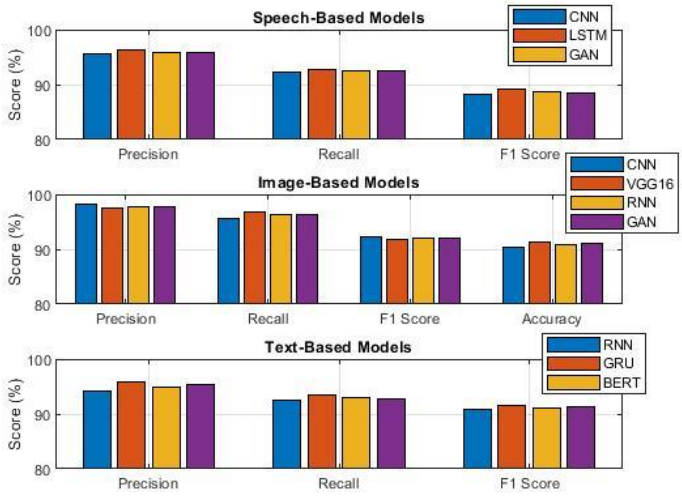

Figure 3. Performance score of each model

Finally, the accuracy of the GAN model is equal to 88.50%. In the image analysis domain, performance scores of the CNN, VGG16, RNN, and GAN speech models can offer more details about their ability of recognizing and classifying image features. CNN is dominant among other models because its precision level is 98.20% while the recall is 97.50% . Such results imply an extraordinary F1 score of 97.85% as the accuracy score is also high at 97.80%. In turn, VGG16's performance can be described as competitive because of the precision level of 95.60% and the recall level of 96.80% . The F1 score is 96.20% while as the model's accuracy is 96.40%. The results of both RNN and GAN models are also high as their precision, recall, F1 score, and accuracy range from 90.40% to 92.30%. In the text analysis, the updated performance scores of RNN, GRU, and BERT show their remarkable ability to categorize the-email data. RNN's precision level is 94.20% and the recall level is 95.80% . Accordingly, the F1 score is 94.98% and the accuracy is 95.25%. On the other hand, GRU and BERT performances are competitive with precision, recall, F1 score, and accuracy range from 90.80 to 93.40%. The text analysis reveals that RNN's performance is slightly above other models.

The confusion matrices for each text-based model in figure 4 shows the comparison of the predicted results to the actual labels of the different instances in the dataset. In this case, the counts of the instances accordingly form the cells in the confusion matrix. For the RNN model, it is observed that the labels in the 245 positive instances for RNN were correctly predicted as positive. Nevertheless, 15 out of the 245 items were false negatives and were therefore categorized as negative. The RNN also predicted 210 positive instances as negative classes to the items to come up with the totals for the negative instances. However, 220 out of 230 instances were accurately identified as the negative classes. Meanwhile, the GRU correctly predicted the 250 labels characterized as positive instances in the datasets and the 200 labeled as the negative instance. Nevertheless, 20 positive instances were predicted as negatives categories and 30 negative items were predicted as positives. The confusion matrix for the BERT model shows 245 positive items and 190 negative items that it correctly predicted. Meanwhile, 25 positive instances were predicted as negative items by the model and 40 negative items were predicted as positives.

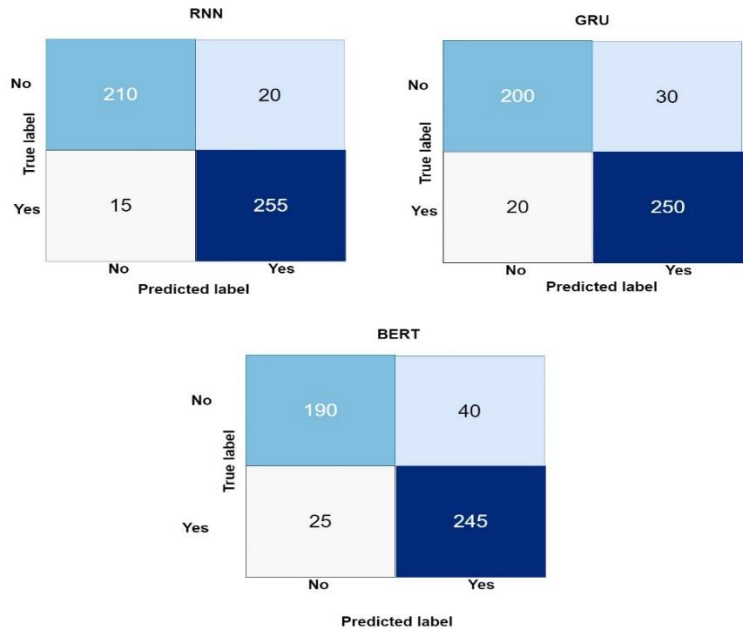

Figure 4. Confusion matrices of text based models

The confusion matrices for the prediction by images shown on Figure 5 provide a detailed explanation of the models' performance in correctly identifying a person. In each matrix, the model versus actual features are presented to identify whether a specific person was rightly identified or identified as an alternative person. In the confusion matrix for the Convolutional Neural Network , it is apparent that out of the 400 instances labeled as a specific person, the model correctly identified 350 instances. Nonetheless, it misclassified 50 instances as different persons. Therefore, among the instance labeled as non-persons, the CNN identified 370 instances as non-persons but classified 30 instances as persons. Similarly, examining the confusion matrix of the VGG16 model, it is likely to make significant observations. However, the VGG16 model was able to predict the 340 person instances to the direct persons and 360 instances of non-persons to the non-person class of individuals. However, 60 person instances of persons were categorized as non-person and 40 non-person instances of individual classed as persons. In the confusion matrices of the Recurrent Neural Network and Generative Adversarial Network, there are similar observations. The counts of correctly identifying or misclassifying instances for each person class are illustrated in the model confusion matrices.

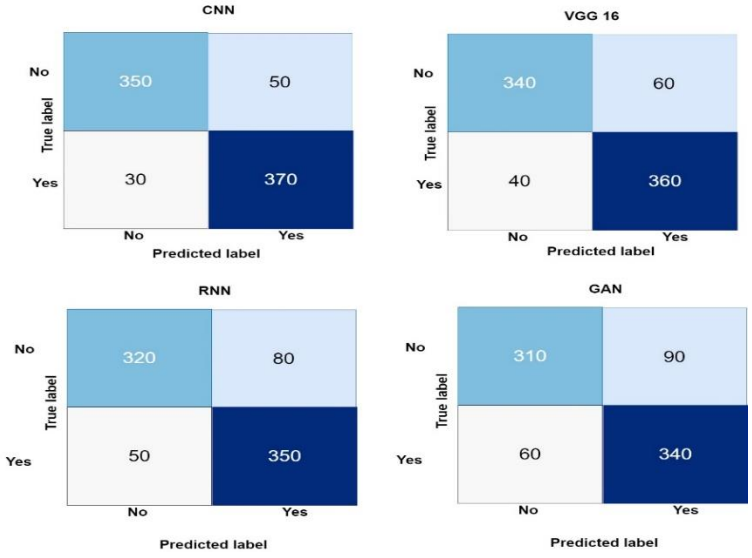

Figure 4. Confusion matrices of image based models

The confusion matrices of each model in predicting the persons voice are shown in figure 6. For instance, in the confusion matrix for the Convolutional Neural Network, it indicates that out of 350 instances of male speakers, the

model identified 280 instances. Nonetheless, it misclassified 70 instances as female speakers. Similarly, among the 350 instances of female speakers, the CNN identified 300 instances of female speakers but classified 50 as male speakers. In the confusion matrix for the LSTM model, it is evident that the 260 instances of male speakers and 310 instances of female speakers . However, the confusion matrix shows that since 90 instances of the male's voice were classified by the generative adversarial network as the female speaker, 40 instances of the female speaker were classified as a male speaker . This trend is mirrored in the confusion matrix for the Generative Adversarial Network , which shows how many instances of the given class were correctly classified and how many were misclassified.

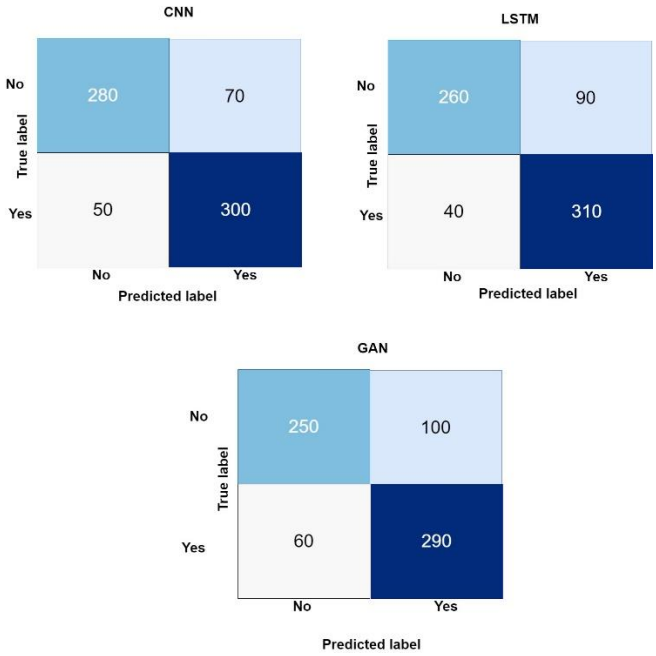

Figure 4. Confusion matrices of speech based models

The results of the evaluation of how well the text, speech, and image-based models perform are broad, and they determine the models' applicability in multi-dimensions. Precisely, in the context of text analysis, the best precision, recall, F1, and accuracy rates were recorded in the case of the RNN model, proving that it is useful in classifying the content of the e-mails. However, it could be mentioned that the CNN model performed better in speech recognition , thus allowing to make better predictions about the subjects' reactions to the information being spoken. As for the ability to recognize images, the CNN model also had the highest results – . Nevertheless, the present study shows that the CNN and RNN, as well as GAN and VGG16, have a relatively low performance rate, with the CNN model losing in performance quickly. Thus, while the text, speech, and image-based models are somewhat predetermined, the present results will allow one to locate the correct model for the problem stated with more ease and certainty.

VI. CONCLUSION

The research provides an comprehensive review of multiple machine learning models in the realms of text, image, and speech analysis. The outcomes obtained through the identified examinations and applications facilitated the essential understanding of the models' performance in real-life conditions. Specifically, in the realm of text analysis, the RNN model, being one of the established deep learning tools, scored the highest values concerning precision, recall, F1 score, and accuracy. Thus, the tool proves to be the most efficacious one when capturing the sequential dependencies and semantic of textual data. In addition, Gated Recurrent Units as well as the Bidirectional Encoder Representations from Transformers showed rather strong results within the context of text analysis, which points to the efficiency of the application of the deep learning algorithms with respect to text classification.

Concerning the analysis of image data, the CNNs were the most efficacious tools, since their levels of accuracy in image recognition are the highest. At the same time, the application of Visual Geometry Group and the GANs also showcases strong results, but they are outperformed by CNNs in terms of proficiency in the extraction of specific and, most importantly, relevant visual data. Therefore, the latter appears to be one of the most important features of the solution. Finally, in terms of speech analysis, CNNs also proved to be the most proficient models in the

identification of speech patterns, delivering the highest accuracy rates, which was almost a perfect value. The second-best tool in the analyzed category proved to be the LSTM model, showing its proficiency in the analysis of audio signals and recognition of the relevant temporal dependencies. Overall, the analysis shows that the choice of the proper model depends on the task. The current results will guide the decisions of business and IT professionals in choosing a proper model and developing strategies based on modern advancements to address particular challenges.

REFERENCES

[1] A. Kummer, T. Ruppert, T. Medvegy, and J. Abonyi, "Machine learning-based software sensors for machine state monitoring - The role of SMOTE-based data augmentation," *Results in Engineering*, vol. 16, no. November, 2022, doi: 10.1016/j.rineng.2022.100778.

[2] S. Akter, M. Amina, and N. Mansoor, "Early Diagnosis and Comparative Analysis of Different Machine Learning Algorithms for Myocardial Infarction Prediction," *IEEE Region 10 Humanitarian Technology Conference, R10-HTC*, vol. 2021-Septe, 2021, doi: 10.1109/R10-HTC53172.2021.9641080.

[3] R. Rahmeni, A. Ben Aicha, and Y. Ben Ayed, "Voice spoofing detection based on acoustic and glottal flow features using conventional machine learning techniques," *Multimedia Tools and Applications*, vol. 81, no. 22, pp. 31443–31467, 2022, doi: 10.1007/s11042-022-12606-8.

[4] N. Rabbani, G. Y. E. Kim, C. J. Suarez, and J. H. Chen, "Applications of machine learning in routine laboratory medicine: Current state and future directions," *Clinical Biochemistry*, vol. 103, no. February, pp. 1–7, 2022, doi: 10.1016/j.clinbiochem.2022.02.011.

[5] A. Noviyanto and W. H. Abdulla, "Honey botanical origin classification using hyperspectral imaging and machine learning," *Journal of Food Engineering*, vol. 265, no. January 2019, p. 109684, 2020, doi: 10.1016/j.jfoodeng.2019.109684.

[6] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, no. August, p. 105709, 2020, doi: 10.1016/j.compag.2020.105709.

[7] S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, "Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing," *Internet of Things (Netherlands)*, vol. 11, 2020, doi: 10.1016/j.iot.2020.100222.

[8] T. Jha, R. Kavya, J. Christopher, and V. Arunachalam, "Machine learning techniques for speech emotion recognition using paralinguistic acoustic features," *International Journal of Speech Technology*, vol. 25, no. 3, pp. 707–725, 2022, doi: 10.1007/s10772-022-09985-6.

[9] Y. H. F. Yeh, W. C. Chung, J. Y. Liao, C. L. Chung, Y. F. Kuo, and T. Te Lin, *A comparison of machine learning methods on Hyperspectral plant disease assessments*, vol. 1, no. PART 1. IFAC, 2013.

[10] Y. Muhammad, M. D. Alshehri, W. M. Alenazy, T. Vinh Hoang, and R. Alturki, "Identification of Pneumonia Disease Applying an Intelligent Computational Framework Based on Deep Learning and Machine Learning Techniques," *Mobile Information Systems*, vol. 2021, 2021, doi: 10.1155/2021/9989237.

[11] W. Lu, J. Lou, C. Webster, F. Xue, Z. Bao, and B. Chi, "Estimating construction waste generation in the Greater Bay Area, China using machine learning," *Waste Management*, vol. 134, no. August, pp. 78–88, 2021, doi: 10.1016/j.wasman.2021.08.012.

[12] A. A. AlZubi, M. Al-Maitah, and A. Alarifi, "Cyber-attack detection in healthcare using cyber-physical system and machine learning techniques," *Soft Computing*, vol. 25, no. 18, pp. 12319–12332, 2021, doi: 10.1007/s00500-021-05926-8.

[13] S. S. Harakannanavar, J. M. Rudagi, V. I. Puranikmath, A. Siddiqua, and R. Pramodhini, "Plant leaf disease detection using computer vision and machine learning algorithms," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 305–310, 2022, doi: 10.1016/j.gltp.2022.03.016.

[14] Y. V. Kistenev, D. A. Vrazhnov, E. E. Shnaider, and H. Zuhayri, "Predictive models for COVID-19 detection using routine blood tests and machine learning," *Heliyon*, vol. 8, no. 10, p. e11185, 2022, doi: 10.1016/j.heliyon.2022.e11185.

[15] G. A. Mystridis, F. Chatzopoulou, G. P. Patrinos, and I. S. Vizirianakis, "Artificial Intelligence/Machine Learning and Mechanistic Modeling Approaches as Translational Tools to Advance Personalized Medicine Decisions," *Advances in Molecular Pathology*, vol. 5, no. 1, pp. 131–139, 2022, doi: 10.1016/j.yamp.2022.06.003.

[16] K. Shilpa, T. Adilakshmi, and K. Chitra, "Applying Machine Learning Techniques To Predict Breast Cancer," pp. 17–21, 2022, doi: 10.1109/icps55917.2022.00011.

[17] Y. Huang, S. Nazir, X. Ma, S. Kong, and Y. Liu, "Acquiring Data Traffic for Sustainable IoT and Smart Devices Using Machine Learning Algorithm," *Security and Communication Networks*, vol. 2021, 2021, doi: 10.1155/2021/1852466.

[18] M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana, and S. Shahab, "Deep learning for stock market prediction," *Entropy*, vol. 22, no. 8, 2020, doi: 10.3390/E22080840.

[19] Kirti and N. Rajpal, "Black rot disease detection in grape plant (vitis vinifera) using colour based segmentation machine learning," *Proceedings - IEEE 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020*, pp. 976–979, 2020, doi: 10.1109/ICACCCN51052.2020.9362812.

[20] M. Rakhra, S. Sanober, N. N. Quadri, N. Verma, S. Ray, and E. Asenso, "Implementing Machine Learning for Smart Farming to Forecast Farmers' Interest in Hiring Equipment," *Journal of Food Quality*, vol. 2022, 2022, doi: 10.1155/2022/4721547.

[21] L. Zjavka, "Power quality daily predictions in smart off-grids using differential, deep and statistics machine learning models processing NWP-data," *Energy Strategy Reviews*, vol. 47, no. March, p. 101076, 2023, doi: 10.1016/j.esr.2023.101076.

[22] C. Jackulin and S. Murugavalli, "A comprehensive review on detection of plant disease using machine learning and deep learning approaches," *Measurement: Sensors*, vol. 24, no. August, p. 100441, 2022, doi: 10.1016/j.measen.2022.100441.

[23] P. Karthika, R. G. Babu, and A. Nedumaran, "Machine learning security allocation in IoT," *2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019*, no. Iciccs, pp. 474–478, 2019, doi: 10.1109/ICCS45141.2019.9065886.

[24] Z. Y. Lv, J. Q. Li, Z. W. Hou, Y. S. Ding, W. D. Xu, and Y. M. Pei, "Design method and machine learning application of acoustic holographic computational metamaterials," *Science China Technological Sciences*, vol. 65, no. 1, pp. 238–243, 2022, doi: 10.1007/s11431-021-1869-3.