National College of Ireland

# National College of Ireland

## Project Submission Sheet

| | | | |
|---|---|---|---|
| **Student Name:** | Vipin Sharma | | |
| **Student ID:** | x22207406 | | |
| **Programme:** | MSc. Data Analytics | **Year:** | 2023-2024 |
| **Module:** | Data Mining Machine Learning 2 | | |
| **Lecturer:** | Prof. Anu Sahni | | |
| **Submission Due Date:** | 19-05-2024 | | |
| **Project Title:** | Data Mining Machine Learning 2 TABA | | |
| **Word Count:** | 5140 | | |

**I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.**
**<u>ALL</u> internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.**

| | |
|---|---|
| **Signature:** | Vipin Sharma |
| **Date:** | 19-05-2024 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# AI Acknowledgement Supplement

## [Insert Module Name]

## [Insert Title of your assignment]

| Your Name/Student Number | Course | Date |
|---|---|---|
| | | |

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

## AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

| Tool Name | Brief Description | Link to tool |
|---|---|---|
| | | |
| | | |

## Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

| [Insert Tool Name] | |
|---|---|
| [Insert Description of use] | |
| [Insert Sample prompt] | [Insert Sample response] |

## Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

## Additional Evidence:

[Place evidence here]

## Additional Evidence:

[Place evidence here]

# DATA MINING AND MACHINE LEARNING 2 TABA

Vipin Sharma
x22207406@student.ncirl.ie
MSc in Data Analytics
National College of Ireland

## Question No.1 - Case Study
## Topic : Generation of a short video based on a description provided as a paragraph of text

### Introduction

Multimedia content is becoming more and more in demand in the current digital age. Across many different kinds of platforms, short videos in particular have become an important component of entertainment and communication. However, it can take a lot of time and resources to create videos from scratch. Machine learning approaches present effective methods to overcome such challenges by automating the generation of videos from textual descriptions. The ability to create videos from text has major challenge for many different industries. The important part of this project is that he production and consumption of multimedia content can be completely changed through automated video generation, changing everything from marketing and education to entertainment and education. We may improve productivity and creativity in video production by using different machine learning algorithms and deep learning algorithms. Presenting a methodology for various parts of the video generation project is the main goal of this report. Ethical issues, dimensionality reduction, feature engineering, hyperparameter optimization, model evaluation, exploratory data analysis, and modeling technique selection are all discussed. Based on current research in the field, each section will discuss the suggested methodology, explanation, and important factors.

### Structure of the Report

The report is divided into eight main sections, each of section focused on different part of the methodology. The sections are followed as exploratory data analysis, data cleaning, feature engineering, modeling technique selection, hyperparameter optimization, model evaluation, scalability concerns, and ethical considerations. These sections give a complete structure for text-to-video conversion, assumptions, suggested methods, reasons, and other main aspects required to analyse and make the project.

#### Exploratory Data Analysis / Data Cleaning

The Process of Creating Videos with Text Descriptions To start, Everyone knows that exploratory data analysis (EDA) or Data cleaning will help identify potential challenges and opportunities by using EDA to get insightful 9information about the dataset's structure and quality. Missing values, outliers and inconsistency are some of the issues that require resolving through data cleaning for a more dependable dataset. A thoughtful approach makes video generation results more accurate and efficient in terms of model performance and overall quality of the data set. In many papers on EDA and Data cleaning were observed variety types approaches and methodologies used in several papers. Balaji et al. (2019) use two datasets which are found from YouTube and the Kinetics human action video dataset but although complete paper lacked details about data cleaning and preprocessing stage. Similarly, In Wang et al. (2023) the author conducted EDA or data cleaning for provided datasets, identifying patterns, anomalies or relationships between variables; handling missing or inconsistent data for subsequent analysis is important as well. Similarly, In the in Li et al. (2018), the authors used EDA to understand the structure and

quality of the data, which could involve techniques such as data cleaning that address inconsistencies, missing values and outliers. In this paper, the author presented EDA and Data Cleaning through dataset preprocessing involving frame extraction and reshaping with a particular focus on practicality associated with these techniques to clean up data for better model filling. Furthermore, an example is given in Zhao et al. (2024) where there was a proper application of EDA on audio, text and video datasets thereby helping in identifying data structure, distribution and correlations that forms basis for informed choice of e.g., data cleaning methods. In some papers writers don't do the data cleaning or EDA but they discuss them how important are they like in Ansari et al. (2024) where there isn't explicit mention of EDA though the author told that data cleaning was done to know what was happening in the problem domain and also for datasets in order for one to get more insights thus suggesting an EDA section. Similarly, the VidTIMIT dataset is employed by author of Zhang et al. (2022) throughout his analysis without mentioning anything concerning EDA or the passage through which he went while analyzing his own dataset. By reading the contents of these papers on EDA, authors Hu et al. (2022)Wang et al. (2023)Li et al. (2018) suggest that the basic EDA and Data cleaning were not enough to get good results, they have already been used in the study and the authors also mentioned that what are the advanced data cleaning techniques and what were some important library that this study can use for the further study. From the point of reference, this study uses some advanced techniques such as handling outliers, correcting errors, and adjusting the imbalanced data. Also, uses the important libraries mentioned in previous studies like numpy, pandas, matplotlib for efficient data cleaning steps. The important things mentioned by the authors were firstly, it is important to understand that one must conduct a thorough EDA to see the data's characteristics, identify issues that may arise, and inform analysis that will be done later. Secondly, there are many different ways of handling missing values or inconsistencies and outliers in data and documenting them as this will aid in reproducibility. By taking care of these important points create and generate good results with high accuracy.

**Dimensionality Reduction/ Feature Selection**

In the process of creating brief videos from textual descriptions, dimension reduction and feature selection are very important steps. Effectively handling large datasets by using the different dimensionality reduction techniques like Principal Component Analysis (PCA) or t-SNE are used. Dimensionality reduction approaches will affect dimensionality and improve the computational efficiency of subsequent modeling steps by removing the number of features while maintaining the necessary information. To improve the understanding of the model, decrease overfitting, and increase overall performance, feature selection techniques also make it possible to identify and take the most important features for the creation of videos. Selecting the most informative features helps ensure that the generated videos accurately capture the words of the input text while minimizing noise and removing irrelevant information from the text. The dimensionality reduction and feature selection are crucial preprocessing stages that optimize computational resources, speed up the modeling process, and make it easier to produce good videos from the given textual descriptions.

Different papers highlight several methods and approaches that can be applied in the Dimensionality Reduction and Feature Selection. It is a bidirectional transformer architecture that utilizes attention mechanisms to select important input features thereby implicitly reducing dimensionality, which has been used in Hu et al. (2022). Even though explicit dimensionality reduction techniques are not mentioned, the attention mechanism performs dimensionally on textual data. Wang et al. (2023) explicitly applies feature selection techniques like Recursive Feature Elimination (RFE) and dimensionality reduction techniques such as Principal Component Analysis (PCA) for better model performance and less overfitting. In the study.Li et al. (2018) shows the significance of recursive feature elimination or Boruta which ensures that chosen features are pertinent because there may be utilization of PCA reduction by employing feature selection methods involved. For instance, in Kumar et al. (2022) hybrid VAE-GAN frameworked shall allow one to perform dimensionality reduction and feature selection; however more information about their impact can lead to further improvement on knowledge in this context. An efficient way of processing video and audio data for TAgVM model by using PCA, t-SNE and feature selection by mutual information (MIM) can be found in the Zhao et al. (2024). In addition to utilizing pre-trained feature extractors, other approaches such as PCA or t-SNE are proposed in Yang et al. (2018) to enhance visualization and performance of the model. In Zhang et al. (2022) a phoneme-pose dictionary is used to reduce dimensionality. However, it is recommended that different methods including PCA or t-SNE be studied concerning comparative analysis. There are many authorsBalaji et al. (2019) Hu et al. (2022) Ansari et al. (2024) Kim et al. (2020) who are not used any Dimensionality techniques but they mention

how this step is important to select only the important features and suggested techniques that are used for Dimensionality Reduction like PCA, t-SNE, Reduce the size of complex data, used mechanism to learn the important features. For this study, used the same dimensional technique mentioned above like PCA, t-SNE, and VAE-GAN because some of the authors used similar types of feature selection techniques giving good results and increasing interpretability, enhancing model performance, and making high-dimensional data processing more efficient.

**Feature Engineering/ Feature Extraction**

When creating short videos from text descriptions, feature engineering, and feature selection are important components. Feature engineering indicates the process of creating meaningful information from the raw text data, which may give insightful information and suggestions for creating videos. Feature engineering helps create video content that closely matches the content and context of the input text description by extracting important characteristics from the text, such as sentiment, semantic information, or key concepts. A comprehensive reading of several papers reveals different methods for feature engineering and feature extraction. In the paper Köksal et al. (2023) the author extracts hand mask for each fame of EPIC-Kitchens-55 using pre-trained hand segmentation model, extraction, and annotation enhancement for video analysis purposes. In addition, the Kumar et al. (2022) uses Stanford CoreNLP toolbox to extract features such as part-of-speech tags and sentiment analysis from text data. Similarly, meaningful features are extracted from text and audio data using pre-trained models like Clip or Wave2Vec 2.0 in Zhao et al. (2024). Furthermore, GANs are suggested by Raja et al. (2023) as a solution to the problem of text-to-video tasks with advanced feature extraction techniques in mind. Additionally, the authors also employed pre-trained feature extractors for text and video data through their research work on Yang et al. (2018). The paper titled "Phoneme and pose extraction from Audio and Video Data" investigated phoneme recognition systems (PR) capable of lipreading while watching lip motion Zhang et al. (2022) also the author does not use but mentioned using mutual information or recursive feature elimination to get the good results. Lastly, skip thought vectors network along with PCA is applied to text encoding as well as dimensional contraction in Kim et al. (2020) but the author also mentioned using the more important features engineering techniques such as word embeddings or attention mechanisms used to improve the text to video generation performance. After reviewing all these papers, Wang et al. (2023)Ansari et al. (2024) suggested Important Techniques for feature engineering and feature extraction such as random forest, XGBoost, or some deep learning models to create new features from the existing ones and the mutual information and recursive features elimination is also used to identify the most relevant features. Applying these techniques for further study to improve the results of the previous study and enhance the model's important features for training purpose.

**Choice of Modelling Techniques**

The modeling techniques plays a very important role in the generation of short videos based on textual data, which improve the quality, efficiency, and effectiveness of the video generation process. Selecting the appropriate modeling techniques is very important and it must be suitable for our algorithms, architectures, and frameworks to translate textual descriptions into contextually relevant video content. Different modeling techniques have their own advantages and disadvantages. To create the model which can use generate the video based on the text there are many traditional machine learning algorithms and pre-trained deep learning modeling. but For this research, the most popular model, generative adversarial networks (GANs) are used because in the past few times, it have gained popularity for their ability to produce realistic and high-quality video content by learning from textual descriptions. Similarly, recurrent neural networks (RNNs) and transformer-based models are also used to generate a well-suited video for the given text description. From the review of many papers, it can be seen that in text-to-video synthesis numerous models are used In Hu et al. (2022), a bidirectional transformer architecture is used which helps a multi-modal framework that combines two things text encoder and video decoder both are used to capture long-range text effectively. Wang et al. (2023), on the other hand, argues for suitable machine learning algorithms chosen to handle specific tasks rather than sticking to traditional methods such as logistic regression or neural networks. In addition, Li et al. (2018) uses the ensembling techniques and combines them with cross-validation to enhance model performance. The approach of paper 6 introduces an adaptable video generation style with motion estimation layers and GANs incorporated into its framework. Additionally Ansari et al. (2024) relies on GANs and VQ-VAE-2 for text-to-image and video synthesis, indicating that there are other options including transformers also available for investigation. Furthermore, Kumar et al. (2022) employs hybrid VAE-GAN as well as video generating

method which could perhaps offer more insight when looking at the choice of methodology required deepening understanding in the field. Moreover Zhao et al. (2024) has shown that novel structures such as the collaborative text and audio-guided video modifier along with text-guided video generator can assist in solving this problem efficiently. Similarly conditional GANs can be used to generate videos from texts while considering various types of GANs as suggested by Yang et al. (2018); multimodal recurrent architecture for text-to-video synthesis, with possibilities for exploring transformer-based models or graph neural networks. In Zhang et al. (2022), Kim et al. (2020), and Mazaheri & Shah (2022) GANs are used for video generation from given text. Mostly all the models used by all the previous researchers so there is nothing new to apply but by reading all the papers found that some authors Ansari et al. (2024)Kumar et al. (2022) use the various GAN models and deep learning models but do not perform the compare the performance of the models and also not give the more details about the model which is used for the study and what are the advantage and disadvantage of the used model. When applying any modeling techniques, make sure to mention everything about the models like advantages and disadvantages and if using more than one model always compare the model performance and accuracy.

**Hyperparameter Optimisation**

Creating a Model for video generation based on textual description is easy but performing the Hyperparameter optimization affects the model's performance, and accuracy. The Hyperparameter is an important step when working with machine learning and deep learning models. If we talk about video generation, hyperparameters control various things which is important to understand the words of a given long text such as such as learning rate, batch size, regularization strength, and network architecture parameters. Optimizing these parameters maximizes the model's ability to learn meaningful words from the textual description and generate high-quality videos while minimizing some issues like overfitting and underfitting. To learn more read various papers where the hyperparameter optimization for text-to-video synthesis and video generation tasks are implemented. In Hu et al. (2022), the author used a pre-trained transformer model for his text encoder and a GAN system to decode videos and fine-tune them using movie datasets without talking about hyperparameter optimization techniques. Li et al. (2018) proposes possible ways of tuning models through hyperparameter optimization such as grid search, random search, or Bayesian optimization. Conversely, Zhao et al. (2024) optimizes hyperparameters for their TAgVM model by use of grid search and cross-validation and guarantees generalization capabilities. However, both Raja et al. (2023) and "Text2viedo" lack mention of any hyperparameter optimization technique thus suggesting that there might have been an oversight in the refinement of the performance of these models. In this case, Kim et al. (2020) does a given number of iterations in both text-to-image and evolutionary generation sections thereby meaning hyperparameter tuning. By reviewing the literature review found that how parameter tuning is important and affects the accuracy of the model also we found that the paper Li et al. (2018)Balaji et al. (2019)Ansari et al. (2024)Kumar et al. (2022)Mazaheri & Shah (2022) not performing the hyperparameter tuning. There are various ways mentioned above used for doing the parameter tuning. Also, there are some advanced techniques used for hyperparameter tuning to improve the accuracy of the previous models such as Grid search, Random search, Bayesian optimization, and evolutionary algorithm. Many authors did not use the cross-validation techniques implementing the cross-validation techniques can further validate the effectiveness of the selected hyperparameters to solve the previous issue. Also by integrating valuable techniques into their methods which can be used to improve the performance of a model after reviewing these insights on hyperparameter optimization.

**Model Evaluation**

Working with any project based on machine learning or deep learning the Model evaluation part is important because it will calculate the accuracy, consistency, and overall quality of the model which helps understand the model behavior. When creating a short video from the text data it is very important that the model capture the text properly and gives good accuracy without showing overfitting and underfitting. There are various statistical metrics are used to evaluate the model such as mean squared error (MSE), mean absolute error (MAE), Precision, recall, and Frechet Inception Distance (FID) with qualitative evaluations, such as user feedback and human perceptual studies. Evaluating the models on different statistical metrics with different datasets gives the guarantee that the models are ready to use in a real-world scenario. While conducting the literature review of the paper I saw that the authors used different types of methodologies and metrics. In the Balaji et al. (2019) and Li et al. (2018) author uses classification accuracy which includes recall, precision, and F1-score over specific metrics relating to the

respective tasks. On the other hand, Wang et al. (2023) highlights the importance of metrics such as accuracy, precision, recall, and F1 score where it proposes methods like cross-validation for robustness. In addition, Li et al. (2018) promotes advanced evaluation techniques such as cross-validation for the robust performance of the model. In Köksal et al. (2023), it can be seen using Fréchet Inception Distance (FID), Fréchet Video Distance (FVD), and Learned Perceptual Image Patch Similarity (LPIPS) metrics for evaluating video generation that finds variation in metric choice. Another way of achieving a detailed understanding of model performance involves additional metrics like FID, PSNR, or SSIM that were used in Ansari et al. (2024). Next, In Zhao et al. (2024) the author provides an extensive evaluation through metrics like Mean Squared Error (MSE), Inception Score (IS), Frechet Inception Distance(FID), and comparison with state-of-the-art models. Similarly, Zhang et al. (2022) does a different way of evaluation by taking user feedback when checking the quality of a model. Further on still Kim et al. (2020) and Mazaheri & Shah (2022) use FID and IS for evaluating the generated images and videos for improved evaluation. All Evolution metrics are used in the above-mentioned papers but the most important metrics are MSE( Mean Squared Error), classification metrics, and MAE (Mean Absolute Error). There are also some advanced metrics that we can use for evaluating the model such as F1 Score, Area under the Curve (AUC), PSNR, and Bootstrapping. These all help to evaluate the model perfectly without causing any problems and understanding of the models' performance.

**Scalability Issues**

Scalability issues play a crucial role in the generation of short videos based on text data, especially when a large dataset is involved, and deploying the model is in real-world scenarios. When the dataset size increases or the complexity of the model increases there is a requirement of memory requirement, processing speed of the machine, and scalability of the algorithm. If talk about short video generation there are various types of scalability issues occurring just as the size and resolution of the video, large datasets that are used for building the model, deploying the model requiring large computational power machines, and resource location. It is very important to address these problems across all the 14 papers that I studied mostly the two scalability issues the first is related to model deployment and the second is handling large datasets. In the paper Li et al. (2018) and Ansari et al. (2024) authors have discussed the importance of deploying the model in a production environment and discussing potential applications across industries. In Zhao et al. (2024) author faces the scalability issue due to the high-resolution and large-scale nature of datasets and resolves this issue with multi-scale training strategies. In another Zhang et al. (2022) addresses that author addressing computational resource requirements and strategies for scaling up the model for the deployment. In the end, I conclude that by reviewing various papers the two most common problems are dataset size and computational resource requirements which are important factors for the scalability issues always take care of these two things. There are various suggested techniques used for solving this issue while working with the project like distributed computing techniques such as Apache Spark, Hadoop, and Docker.

**Ethical implications**

Ethical implications are very important when we work with image datasets, textual datasets, and creating a video from text data. In this, we make sure that we are fair and very respectful if any human data or public information is used. Always take permission if we use the private dataset for our study so that we protect the people's privacy. One more important thing is that when creating a video from the text it makes sure that the video which is generated does not spread wrong information. By thinking about these things carefully, we can make sure our videos are trustworthy and respectful to everyone. In the previous papers, authors make sure they consider ethical implications across the papers while some of them address ethical concerns. In Wang et al. (2023), methods such as secure privacy data analysis and explainable machine learning techniques are suggested, but the discussion is mainly on evaluating fairness, bias, privacy, and interpretability to safeguard the ethical integrity of models. Similarly, Li et al. (2018) suggests that biases from data or possible misuse of models cannot be ignored by authors. On the other hand; Ansari et al. (2024) just mentions accessibility and inclusivity in passing. In comparison, "Paper-9" prioritizes moral principles more with an emphasis on minimizing opinions, showing respect for other's rights, and guarding privacy.

# Question No.2 - Paper Review
# Topic: Voice spoofing detection for multiclass attack classification using deep learning

### Structure Title

The title of the paper Voice spoofing detection for multiclass attack classification using deep Learning clearly describes the complete article. The article includes information on various deep learning models like Convolutional Neural Networks (CNN), WaveNet, and recurrent neural network (RNN) variants such as Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) models used for the analysis to classify the binary and multi-class attacks. The content of the article well clear and structured with all sections having proper headings and subheadings, papers are also cited in the Harward referencing style there are very few small formatting mistakes that can easily fixed.

### Abstract

Yes, the abstract section provides a small summary of the article. It describes the main issues of voice spoofing detection and How the suspect takes advantage of Voice spoofing, also talks about the classification of real and fake audio by using the different pre-trained deep learning models. The evaluation of the model used the large dataset which includes 419426 audio files and for the development to real-time classification the user will upload the audio files and microphone voice audio also described in the abstract section. Also, there are performance metrics like the False positive rate (FPR) of all the deep learning models are also mentioned which is good for the readers. These all are important points that are important for any abstract section.

### Introduction

Yes, the introduction part describes the issues properly, and the author's goals and the voice spoofing issue are accurately and stated in the paper introduction section. There are six important key features robust dataset, feature evaluation, binary and multiclass classification, Model Deployment Testbed, and real-time classification will become the basis for the author's effort to develop the best voice spoofing model that classifies the audio. Background knowledge of Voice spoofing is also mentioned Voice spoofing attacks are common nowadays as technology becomes more and more advanced and integrated into our daily lives. Today every smartphones, laptop, and tablets work with voice assistant technology, It will be very easy for someone to access all the important details of the laptop using spoofing voice. The spoofing voice is a way to create, manipulate, and convert the real voice using different technology. Recorded voice can be used to control devices connected to the internet such as smartphones, laptops or In some cases it is also used to authenticate access to accounts through automated recognition systems biometrically. By doing these kinds of things payment fraud, property fraud anything can happen and it will result in a large loss. The author uses the different types of pre-trained deep learning classifier models such as Convolutional Neural Networks (CNN), WaveNet, and recurrent neural network (RNN) variants such as Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) are used to classify the different type of audio categories real, converted, synthesized, and replay to prevent the voice spoofing attacks.

### Highlights

- Using a larger dataset includes 419426 audio files larger than earlier research.
- False Positive rate of 0.0216 for binary classification and 0.003 for multiclass.
- Deep learning methods classify real and fake audio along with attack vectors.
- Four different pre-trained deep learning models are used to make the model.
- Different types of audio categories are real, converted, synthesized, and replay.

### Methodology

The paper utilizes a large dataset which is a collection of five multiple datasets like VSDC, ASVSpoof2019, ReMASC, ASVSpoof2017, and ASVSpoof2015 and every dataset uses a different variety of hardware tools

like a microphone to capture real-world spoken audio or phone recordings or a speaker for reply attack audio shown in Table 2. The Dataset distribution in this paper is clearly shown in Table 3 with 4 classes Real audio contains 51009 audio, Synthsised audio contains 99312, Converted audio contains 59794, and finally, the last class Replay audio has a total of 209310 audio. Yes, the design is good for answering the research question. The author uses the various types of deep learning algorithms to look into spoofing detection problems as binary and multiclass classification problems because as we also know the deep learning model works better than machine learning models on image and audio datasets. Users give the inputs via the microphone or pre-recorded dataset via the deployed models' web services.

In the methodology section, everything is defined properly which is used for the research but there are some details like hyperparameter tuning, which are the hyperparameter parameters used to increase the accuracy of the model. Also, they do not provide the specific configuration of the hardware and software used to build the train and test the deep learning models. The coding part is also not available to build the deep learning model that's why I say that it is a little bit impossible to replicate the research.

Yes, the paper follows the procedure and explains everything step by step starting with explaining the data collection process then defining a little about the different deep learning models, and then explaining briefly Audio feature extraction techniques, Normalisation of audio files, and performance metrics. Also, Fig 1 explains the framework for voice spoofing detection and multiclass attack classification using deep learning. All the steps are explained in a meaningful way the first author explains dataset collection, the second, explains the deep learning algorithm, the third step explains audio feature extraction, the fourth normalization of the audio file so that all come at the same size, and then Evaluation and finally Conclusion. The methods are used in the paper are not new because the author uses the deep learning algorithm which now widely used in every sector. Specially for the image and audio classification the deep learning models performed very well. The author gives a very good explanation of deep learning models and how they will used for voice spoofing detection. Nothing was mentioned about the sampling appropriately but the author split the data into 70% for the training and 30% for testing from the 90% of data and 10% of data is using for the validation purpose. The 70(Train)/30(test) ratio is commonly used while working with the machine learning and deep learning tasks. The materials and equipment used in the paper are not described properly but the author mentions some Python library TensorFlow which we used for implementing the deep learning models and Northern Ireland High-Performance Computing Kelvin high-performance computing service is used for training the models. Yes, the paper makes it clear what types of data were recorded there are 5 different types of publically available data used for voice spoofing detection tasks, and Table 4 the author describes the important measurements such as Frequency, Duration, and File format of the audio file.

**Results**

The Paper uses various deep learning models to classify the audio whether it is real or fake. The Flase positive rate (FPR) rate 0.0216 was achieved by the binary class convolutional neural network while on the other side, the multi-class achieved an FPR is 0.003. The multi-class models are used to determine whether the audio is real or fake and also determine on which attack conversion, synthesis, or replay was employed in the spoofed audio. In the paper, the FPR ranges between 0.0231 to 0.4836 for binary classification and 0.0046 to 0.1864 for multiclass classification. It is also shown that the proposed model which is used in the paper has the highest F1-score then the state-of-art-method for both binary and multiclass classification. The results show the effectiveness of the proposed voice spoofing attack model and the appropriate analysis is correct. The using of large size of audio datasets deep learning model achieves high accuracy and low false positive rate on both binary and multiclass classification tasks.

**Conclusion/ Discussion**

Yes, the claims are supported by the results presented in the given paper. The author uses the large dataset to train and testing the model and the model performs very well and give good accuracy. The Pre-trained deep learning model effectively detects voice spoofing attacks with high accuracy and low false positive rate (FPR). The author also compared their results with the previous studies and also discussed the limitations of the previous papers and include those limitations in the paper. Yes, the article supports the previous theories that deep learning models are effective for voice spoofing detection. By using the large variety of datasets applied various deep learning algorithms get good results and the model performed very well with the large dataset. The conclusion section explains how this research has advanced scientific knowledge and also discusses the implication of their research for automated speaker verification systems. It is also mentioned how the model effectively suggested whether the audio is real

or fake by detecting the audio. The future work is to use some other advanced pre-trained deep learning models and expand the study for further analysis.

**Language**

There is no grammatical error found in the paper. All the text is very clear and well-written making it easier for any author to understand the science of presenting. The article presents every section and subsection very clearly so that anyone can understand the science of the paper and go with the flow. The formatting of the paper is changed a little bit according to the figures and tables.

**Previous research**

Yes, the paper references the previous research very correctly. The author read many works related to voice spoofing detection and countermeasures and gives detailed explanations about state-of-the-art methods. After reading the literature review section states that all the important works related to the voice spoofing domain are properly cited the references appear very accurately without any mistakes in the section and all the paper links and citations are available at the end of this paper which includes the name of the author, title of the paper, and In which year it was published.

# References


Ansari, H., Tekade, S., Ghagre, C., Kolte, S. & Shende, T. (2024), 'Text-to-image-and-video generator using machine learning', *International Journal of Innovative Research in Technology and Science* **12**(2), 219–226.

Balaji, Y., Min, M. R., Bai, B., Chellappa, R. & Graf, H. P. (2019), Conditional gan with discriminative filter generation for text-to-video synthesis., *in* 'IJCAI', Vol. 1, p. 2.

Hu, Y., Luo, C. & Chen, Z. (2022), 3make it move: controllable image-to-video generation with text descriptions, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 18219–18228.

Kim, D., Joo, D. & Kim, J. (2020), 'Tivgan: Text to image to video generation with step-by-step evolutionary generator', *IEEE Access* **8**, 153113–153122.

Köksal, A., Ak, K. E., Sun, Y., Rajan, D. & Lim, J. H. (2023), 'Controllable video generation with text-based instructions', *IEEE transactions on multimedia* .

Kumar, S., Ghai, V., Jha, A. & Sharma, S. (2022), Role of artificial intelligence in generating video, *in* '2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)', Vol. 1, IEEE, pp. 697–701.

Li, Y., Min, M., Shen, D., Carlson, D. & Carin, L. (2018), Video generation from text, *in* 'Proceedings of the AAAI conference on artificial intelligence', Vol. 32.

Mazaheri, A. & Shah, M. (2022), Video generation from text employing latent path construction for temporal modeling, *in* '2022 26th International Conference on Pattern Recognition (ICPR)', IEEE, pp. 5010–5016.

Raja, S., Mierudhula, S. & Potheeswari, J. (2023), Text to video generation using deep learning, *in* '2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)', IEEE, pp. 1–7.

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X. & Zhang, S. (2023), 'Modelscope text-to-video technical report', *arXiv preprint arXiv:2308.06571* .

Yang, X., Zhang, T. & Xu, C. (2018), 'Text2video: An end-to-end learning framework for expressing text with videos', *IEEE Transactions on Multimedia* **20**(9), 2360–2370.

Zhang, S., Yuan, J., Liao, M. & Zhang, L. (2022), Text2video: Text-driven talking-head video synthesis with personalized phoneme-pose dictionary, *in* 'ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 2659–2663.

Zhao, M., Wang, W., Chen, T., Zhang, R. & Li, R. (2024), 'Ta2v: Text-audio guided video generation', *IEEE Transactions on Multimedia* .