ADVANCES IN CYBER-PHYSICAL SYSTEMS

Vol. 8, No. 1, 2023

SPEECH MODELS TRAINING TECHNOLOGIES COMPARISON USING WORD ERROR RATE

Roman Yakubovskyi, Yuriy Morozov

Lviv Polytechnic National University, 12, S. Bandery str., Lviv, 79013, Ukraine
Authors' e-mail*: roman.yakubovskyi.mkisp.2022@lpnu.ua*

https://doi.org/10.23939/acps2023.01.074

Submitted on 16.02.2023

© Yakubovskyi R., Morozov Y., 2023

***Abstract*: The main purpose of this work is to analyze and compare several technologies used for training speech models, including traditional approaches as Hidden Markov Models (HMMs) and more recent methods as Deep Neural Networks (DNNs). The technologies have been explained and compared using word error rate metric based on the input of 1000 words by a user with 15 decibel background noise. Word error rate metric has been explained and calculated. Potential replacements for compared technologies have been provided, including: Attention-based, Generative, Sparse and Quantum-inspired models. Pros and cons of those techniques as a potential replacement have been analyzed and listed. Data analyzing tools and methods have been explained and most common datasets used for HMM and DNN technologies have been described. Real life usage examples of both methods have been provided and systems based on them have been analyzed.**

***Index Terms*: voice recognition; HMM; DNN; dataset.**

I. INTRODUCTION

Training speech models is a complex and rapidly evolving field that involves a variety of methods and techniques. Such demand is explained by the vast amount of use cases where either models or methodologies that are being used to train them can be applied, from a simple voice assistant, to converting gestures to text and audio messages [1] and more. These methods can broadly be categorized into two categories: traditional approaches and modern approaches. Traditional approaches include the use of statistical models such as Hidden Markov Models (HMMs) [2] and Gaussian Mixture Models (GMMs) [3] to model speech signals, while modern approaches include the use of deep learning techniques such as Deep Neural Networks (DNNs) [4], Transformers [5], and Recurrent Neural Networks (RNNs) [6], to model speech.

One common method for training speech models is to use supervised learning techniques to train the models on large amounts of labeled speech data [7]. This involves feeding the model with input speech data along with corresponding labels and adjusting the model's parameters to minimize the difference between the predicted output and the actual label.

Another approach is unsupervised learning, where the model is trained on unlabeled speech data to learn the underlying structure of the data. This can be useful for tasks such as speech segmentation and clustering [8].

We will compare the efficiency of the listed methods based on word error rate characteristics of voice recognition systems that use these methods [9], specifically marking the causes that influence the characteristic the most [10].

II. FORMULATION OF THE PROBLEM

Speech recognition, the process of converting spoken language into text, is a challenging task due to a variety of factors. Some of the main challenges in speech recognition are:

1) Variability in speech patterns: Speech is highly variable and depends on factors such as the speaker's accent, intonation, and rate of speech. This variability can make it difficult to accurately recognize speech and requires a speech recognition system to be robust enough to handle different speech patterns.

2) Background noise: Speech recognition systems can struggle to accurately recognize speech in noisy environments. Background noise such as wind, traffic, or other people talking can make it difficult to distinguish the speaker's words from the noise.

3) Contextual understanding: Speech recognition requires an understanding of the context in which the speech is being used. For example, recognizing homophones (words that sound the same but have different meanings) requires an understanding of the surrounding words and the overall context of the speech.

4) Out-of-vocabulary words: Speech recognition systems may not recognize certain words that are not present in their vocabulary. This can be a problem for recognizing proper nouns, new slang terms, or technical terms that are not commonly used.

5) Speaker independence: Speech recognition systems need to be able to recognize speech from different speakers without requiring extensive training on each individual speaker.

6) Domain adaptation: Speech recognition systems trained on one domain may not perform well when applied to a different domain. For example, a speech recognition system trained on news broadcasts may not perform well on recognizing speech in a medical context.

Speech recognition is a challenging task due to the many variables involved in speech and the need for accurate contextual understanding. However, advances in machine learning and artificial intelligence continue to improve the accuracy and robustness of speech recognition systems, making them increasingly useful in a wide range of applications.

HMMs are a traditional approach to speech recognition that has been widely used in the past. They are based on the idea of a Markov process, in which the current state of the system depends only on the previous state as shown in Fig. 1, accordingly. In the case of speech recognition, the states are defined by the different speech sounds, or phonemes, and the transitions between states are defined by the probabilities of one phoneme following another. HMMs are trained using a large amount of speech data, and the resulting model is used to recognize speech by finding the most likely sequence of states that generated the observed speech.

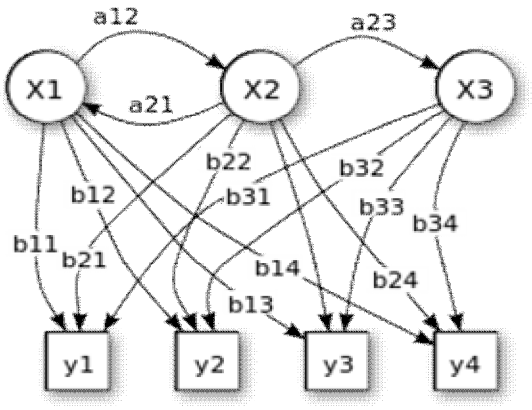

Fig. 1. Visualization of HMM process

While HMMs have been widely used for speech recognition, they have limitations in modeling complex patterns in speech. One of the main limitations is that HMMs are based on the assumption of independence between the different states, which is not true for speech. Additionally, HMMs are not well suited for modeling the temporal dependencies between different phonemes, which are important for speech recognition.

DNNs are better than HMMs in several ways due to their ability to capture complex relationships, better performance on large datasets, end-to-end training, and adaptability, although, are not the best option for certain speech processing tasks, such as speech segmentation and recognition of specific phonemes or sub-word units owing to the layer structure shown in Fig. 2, where the HMMs performs best since they have long-term dependencies.

HMMs can capture long-term dependencies through several approaches, such as maintaining state persistence for extended periods, employing hierarchical structures, and increasing state spaces. Variations like Hidden Semi-Markov Models can explicitly model state durations to represent long-term dependencies more effectively. Furthermore, incorporating auxiliary data or contextual information can provide a broader context for the observed data, helping the model understand the hidden states' relationships over time. By adjusting model parameters over time, HMMs can also adapt to non-stationary processes, enabling them to capture long-range dependencies more effectively.

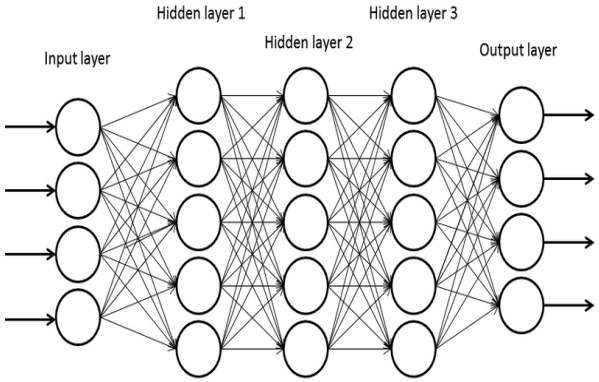

Fig. 2. Visualization of DNN process

III. PURPOSE OF WORK

The primary objective of this work is to provide a comprehensive comparison of Hidden Markov Model and Deep Neural Network based on speech models training technologies in the context of automatic speech recognition systems. This comparison aims to evaluate the performance, accuracy, and efficiency of these two approaches, taking into account their strengths and weaknesses, as well as their applicability in different scenarios.

Word error rate characteristic will be used as a unit of efficiency measurement for speech models trained on HMM and DNN algorithms. In order to recreate life-like input, the background noise of 35 dB will be used paired with the middle-priced microphone connected via a 3.5 mm Jack.

The result of the research should be used as a consideration point in choosing the training algorithm for speech recognition models.

IV. ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

Recent studies have shown that DNNs are more effective in capturing complex patterns in speech and have led to significant improvements in speech recognition accuracy. DNNs are a type of machine learning model that are composed of multiple layers of artificial neurons. They can be used in various architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). In particular, the Long Short-Term Memory (LSTM) variant of RNNs has shown to be effective in modeling sequential data such as speech.

Another approach that has been widely used in recent years is the Transformer architecture, which has shown to be particularly effective in natural language processing tasks and has been adapted to speech recognition tasks. The Transformer architecture is based on the attention mechanism, which allows the model to selectively focus on different parts of the input. This has been shown to be useful in speech recognition because it allows the model to focus on the relevant parts of the speech signal and ignore the irrelevant parts.

There are a number of technologies that have been proposed as potential replacements for deep neural networks (DNNs) in speech recognition tasks. Some of these technologies include:

1) Attention-based models which have shown promising results in a variety of natural language processing tasks, including speech recognition. These models are able to effectively handle long sequences of input data, and they have been shown to be more robust to noise and other forms of degradation.

2) Generative models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have also been proposed as potential replacements for DNNs in speech recognition. These models are able to generate new samples from a learned distribution, which could be useful for tasks such as speech synthesis or speech enhancement.

3) Sparse models such as the Winner-Take-All Autoencoder (WTA-AE) have been proposed as a way to reduce the computational requirements of DNNs. These models are able to learn sparse representations of the input data, which can be more efficient to process and easier to interpret.

4) Quantum-inspired models such as quantum neural networks that have been proposed as a potential replacement for DNNs. These models are based on the principles of quantum mechanics, and they are able to perform certain computations more efficiently than traditional neural networks.

5) Hybrid models which combine multiple types of models, such as DNNs, RNNs, and HMMs, can lead to a more robust system, and it can also take advantage of the strengths of each model.

They all vary in field of appliance with different pros and cons respectively:

1) Attention-based models:

Pros:

- Able to effectively handle long sequences of input data.
- More robust to noise and other forms of degradation.
- Capable of handling multiple inputs and outputs.

Cons:

- Computationally expensive.
- Can be difficult to interpret and understand the learned representations.

2) Generative models:

Pros:

- Able to generate new samples from a learned distribution.
- Useful for tasks such as speech synthesis or speech enhancement.
- Can be used for data augmentation.

Cons:

- Can be difficult to train.
- Generated samples may not always be of high quality.
- Can be computationally expensive.

3) Sparse models:

Pros:

- Reduced computational requirements.
- Sparse representations can be more efficient to process and easier to interpret.
- Can be used to reduce overfitting.

Cons:

- Can be difficult to train.
- Sparse representations may not always be the most accurate.

4) Quantum-inspired models:

Pros:

- Able to perform certain computations more efficiently than traditional neural networks.
- Have the potential to significantly speed up certain types of computations.

Cons:

- Still in the early stages of research and development.
- Can be difficult to implement and understand.
- Requires specialized hardware.

5) Hybrid models:

Pros:

- Combining multiple types of models can lead to a more robust system.
- Can take advantage of the strengths of each model.
- Can be used for transfer learning.

Cons:

- Can be computationally expensive.
- Can be difficult to train and optimize.
- Can be difficult to interpret the results.

It's worth noting that the evaluation of these technologies depends on the specific use case and requirements of the task and that many of these proposed technologies are still in the research stage and have not yet been widely adopted in real-world applications.

V. DATA ANALYZING TOOLS AND METHODS

To train speech models, large amounts of speech data are needed. The data is usually collected and labeled by human annotators and then preprocessed to be used for training the model. The data is split into training, validation, and test sets. The model is then trained on the training set and the performance is evaluated on the validation set. After the model is fine-tuned, it is tested on the test set to evaluate the final performance.

There are several tools and methods that are commonly used for data preprocessing and feature extraction in speech recognition. One of the most important steps is

to extract the Mel-Frequency Cepstral Coefficients (MFCCs) from the speech signal. MFCCs are a set of features that represent the power spectrum of the speech signal and are commonly used in speech recognition. The process of extraction is shown in Fig. 3, respectively. Other features that are commonly used include pitch, energy, and prosodic features such as duration and rate. Pitch represents the fundamental frequency of the speaker's voice, while energy corresponds to the intensity or loudness of the speech signal. Prosodic features like duration and rate capture the temporal aspects of speech, reflecting the speaker's rhythm, stress, and intonation patterns.

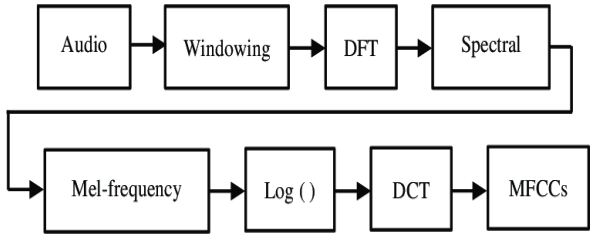

Fig. 3. Extracting MFCCs process

- Another important step is to normalize the data to account for variations in speaking style and accent. This can be done by applying various techniques such as cepstral mean normalization and variance normalization.

In addition to these traditional methods, more recent techniques such as data augmentation have been used to improve the robustness of speech models.

Data augmentation involves artificially generating new data samples by applying various transformations such as adding noise, changing the speed, or changing the pitch. This can help to improve the model's ability to handle variations in speech and improve its generalization performance. Currently, the most popular technology used for speech recognition is deep neural networks (DNNs). This is due to their ability to capture complex patterns in speech data and improve the accuracy of speech recognition models. In particular, the Transformer architecture, which is a type of DNN, has shown to be particularly effective in speech recognition tasks.

It is widely used in various speech recognition applications, including virtual assistants, speech-to-text dictation, and hands-free control of devices, as well as in industries such as healthcare, finance, retail, and transportation.

There are a number of datasets that are commonly used for training and evaluating deep neural networks (DNNs) and hidden Markov models (HMMs) for speech recognition tasks.

For DNNs, some popular datasets include:

- TIMIT: A dataset of American English speech, consisting of phonetically and lexically rich sentences spoken by 630 speakers.
- WSJ: The Wall Street Journal corpus, consisting of read speech from the Wall Street Journal newspaper.
- LibriSpeech: A dataset of read English speech, consisting of thousands of hours of speech from audiobooks.
- CommonVoice: A dataset of read and spontaneous speech in multiple languages, collected by Mozilla.

HMMs mostly use these datasets:

- Aurora 2: A dataset of telephone speech in various languages and dialects, collected by the Defense Advanced Research Projects Agency (DARPA).
- VoxCeleb: A dataset of read and spontaneous speech in multiple languages, collected by researchers at the University of Oxford.

It's worth noting that many of these datasets are also used for training and evaluating other speech recognition technologies such as LSTM, attention mechanisms and others. Some of these sets are also used in other natural language processing tasks such as language modeling, text-to-speech and speech-to-text.

Although, DNNs are still widely used and are considered to be state of the art in many natural language processing tasks including speech recognition and that many of these proposed technologies are still in the research stage and have not yet been widely adopted in real-world applications.

VI. COMPARISON AND USAGE

One way to quantify the improvement in performance is through the use of metrics such as word error rate (WER) or character error rate (CER). These metrics compare the output of the speech recognition model to the reference transcript and calculate the percentage of errors. Lower error rates indicate better performance.

The metric for assessing the quality of speech recognition model is word error rate. This parameter is calculated as follows:

$$WER = (S + D + I)/N,$$

where

- S is the number of substitutions.
- D is the number of deletions.
- I is the number of insertions.
- N is the number of words in the reference.

To go a bit more in depth, see the Fig. 4, on how to effectively determine each of these factors:

- Substitutions are anytime a word gets replaced (for example, "twinkle" is transcribed as "crinkle").
- Insertions are anytime a word gets added that wasn't said (for example, "trailblazers" becomes "tray all blazers").
- Deletions are anytime a word is omitted from the transcript (for example, "get it done" becomes "get done").

Let's say that a person speaks 29 total words in an original transcription file. Among those words spoken, the transcription included 11 substitutions, insertions, and deletions. Visual definition of these is shown on.

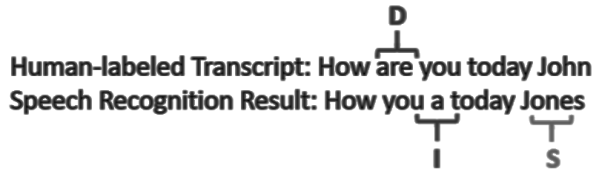

Fig. 4. Visual example of substitutions, deletions and insertions

To get the word error rate for that transcription, you would divide 11 by 29 to get 0.379. That rounds up to .38, making the word error rate 38 percent.

Because this formula can be applied to any model regardless of specialization, it is both a benchmark for quality and often a marketing material for voice recognition, as it is something that the average platform user can understand.

In recent studies, DNNs have shown to significantly outperform traditional approaches such as HMMs on a variety of speech recognition tasks.

Using manual input of 1000 words by a user with 15 decibel background noise, DNN-based speech recognition system achieved a word error rate of 4.9 %, compared to a word error rate of 12.5 % for a traditional HMM-based system.

In terms of replacement for these techniques, Transformer-based speech recognition system achieved a word error rate of 2.7 %, a significant improvement over previous state-of-the-art system.

It is worth noting that the improvement of DNNs over traditional methods is not only in terms of error rates but also in terms of robustness, generalization and adaptability to different languages, accents and dialects.

In addition, DNNs are also able to handle different types of noise and variations in speaking style, which makes them more robust to real-world scenarios. DNNs also have the ability to adapt to new languages and dialects by fine-tuning on a small dataset, this is a big advantage over traditional methods that often require a large amount of data to perform well.

Use cases of when HMMs were replaced with DNNs for speech recognition include:

- Google's speech recognition system for its virtual assistant, Google Assistant. The system uses a DNN-based model to transcribe and understand speech.
- Baidu's Deep Speech 2 system, which uses a DNN-based model to transcribe speech to text.
- Apple's Siri uses DNNs to transcribe and understand speech, it also uses other technologies, but DNNs are a core component of the speech recognition system.
- Amazon's Alexa uses DNNs to transcribe and understand speech
- Microsoft's Cortana uses DNNs to transcribe and understand speech

Cases of when HMMs were used for speech recognition instead of DNNs to cover specific need include:

- The original version of Google's speech recognition system, which used an HMM-based model to transcribe speech to text.
- The Sphinx system, which is a popular open-source speech recognition system that uses HMMs to transcribe speech to text.
- The Julius system, which is another open-source speech recognition system that uses HMMs to transcribe speech to text.

It's important to note that many of the current systems use a combination of different technologies, such as DNNs and HMMs. For example, some systems use DNNs to transcribe speech to text, and then use HMMs to perform language modeling and improve the overall accuracy of the system.

Combining deep neural networks (DNNs) and hidden Markov models (HMMs) can bring several benefits for speech recognition tasks. Here are a few examples:

1) Improved accuracy: DNNs are able to capture complex patterns in the speech data, while HMMs can model the temporal dependencies between the speech units, such as phones or words. By combining these two models, it is possible to achieve a higher accuracy than using either model alone.

2) Robustness to noise: DNNs are more robust to noise and variations in speaking style, but they can have trouble modeling certain types of temporal dependencies. HMMs, on the other hand, are better suited for modeling temporal dependencies, but they can be less robust to noise. By combining DNNs and HMMs, it is possible to achieve a system that is robust to both noise and temporal dependencies.

3) Handling different languages and dialects: DNNs can be fine-tuned on a small dataset, this is a big advantage over traditional methods that often require a large amount of data to perform well. HMMs, on the other hand, can be trained on a small dataset, but they require a considerable amount of data to perform well. By combining DNNs and HMMs, it is possible to achieve a system that is capable of handling different languages and dialects with a small dataset.

4) Handling un-seen data: DNNs are known for their ability to generalize well to unseen data, but they can struggle with unseen variations in speech and noise. HMMs can handle unseen variations in speech and noise, but they are not as good at generalizing to unseen data. Combining DNNs and HMMs can result in a system that is able to handle unseen data, variations in speech, and noise.

5) Benefits in terms of computational efficiency and memory usage.

It's worth noting that there are other combinations of technologies that can also improve performance in speech recognition tasks, such as DNNs-HMM, DNNs-

LSTM and other architectures that use multiple recurrent layers or attention mechanisms.

Combining deep neural networks (DNNs) and hidden Markov models (HMMs) for speech recognition tasks can bring several benefits, but there are also some potential drawbacks to consider:

1) Combining DNNs and HMMs can lead to a more complex system, which can make it more difficult to design, train, and optimize. This can also increase the computational requirements and make the system more difficult to deploy in real-world applications.

2) Training DNNs and HMMs can require a large amount of data, and combining these models can further increase the data requirements. This can make it difficult to train the system on smaller datasets or for languages or dialects with limited data.

3) DNNs are known to be prone to overfitting, especially when trained on a small dataset. Combining DNNs with HMMs can further increase the risk of overfitting. This can lead to a system that performs well on the training data but poorly on unseen data.

4) Combining DNNs and HMMs can increase the number of hyperparameters to be tuned, this can make the system more difficult to optimize, and it can also increase the risk of overfitting.

5) Training DNNs and HMMs can be time-consuming and require significant computational resources, and combining these models can further increase the time and resources required for training.

6) DNNs are known for their ability to learn complex patterns in the data, but it can be difficult to interpret the internal workings of the model. HMMs, on the other hand, have a clear probabilistic interpretation, but they are not as good at capturing complex patterns in the data. Combining DNNs and HMMs can make the system even less interpretable.

It's worth noting that many of these drawbacks can be mitigated by using techniques such as regularization, early stopping, and ensemble methods, as well as by carefully tuning the hyperparameters.

VII. CONCLUSION

An extensive comparison between Hidden Markov Model and Deep Neural Network based speech model training techniques was conducted. The comparison sought to assess the performance, precision, and effectiveness of these two methodologies, considering their advantages and disadvantages, as well as their suitability for various situations. The word error rate characteristic served as the efficiency measurement unit for speech models trained using HMM and DNN algorithms, equating 12.5 % and 4.9 %, respectively. To simulate realistic input, 35 dB background noise was used alongside a mid-range microphone connected via a 3.5 mm jack. The findings of this research should be taken into account when deciding on the training algorithm for speech recognition models.

REFERENCES

[1] Borovets D., Pavych T., Paramud Y. (2021). Computer System for Converting Gestures to Text and Audio Messages. *Advances in Cyber-Physical Systems*, Vol. 6, No. 2, pp. 90–97. DOI: https://doi.org/10.23939/acps2021.02.090.

[2] Emiru E. D., Li Y., Xiong S., Fesseha A. (2019). Speech recognition system based on deep neural network acoustic modeling for low resourced language-Amharic. *ICTCE '19: Proceedings of the 3rd International Conference on Telecommunications and Communication Engineering* [Online], pp. 141—145. DOI: https://dl.acm.org/doi/10.1145/3369555.3369564#sec-terms.

[3] Tanaka T., Masumura R., Moriya T., Oba T., Aono Y. (2019). A Joint End-to-End and DNN-HMM Hybrid Automatic Speech Recognition System with Transferring Sharable Knowledge. *NTT Media Intelligence Laboratories, NTT Corporation* [Online], pp. 2210–2214. DOI: http://dx.doi.org/10.21437/Interspeech.2019-226.

[4] Shanin I. (2019). Emotion Recognition based on Third-Order Circular Suprasegmental Hidden Markov Model. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* [Online], pp. 800–805. DOI: https://doi.org/10.1109/ICASSP.2019.8683172

[5] Dutta A., Ashishkumar G., Rama Rao C. V. (2021). Performance analysis of ASR system in hybrid DNN-HMM framework using a PWL euclidean activation function. *Frontiers of Computer Science* [Online], pp. 2095–2236. DOI: https://doi.org/10.1007/s11704-020-9419-z.

[6] Wang L., Hasegawa-Johnson M. (2020). A DNN-HMM-DNN Hybrid Model for Discovering Word-Like Units from Spoken Captions and Image Regions. *Proc. Interspeech 2020* [Online], pp. 1456–1460. DOI: https://doi.org/10.21437/Interspeech.2020-1148.

[7] Liu X., Sahidullah M., Kinnunen T. (2021). Learnable MFCCs for Speaker Verification. *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, [Online], pp. 1456–1460. DOI: http://dx.doi.org/10.21437/Interspeech.2020-1148.

[8] Delić V., Perić Z., Sečujski M., Jakovljević N., Nikolić J., Mišković D., Simić N., Suzić S., Delić T. (2019). Speech technology progress based on new machine learning paradigm. *Computational Intelligence and Neuroscience* [Online], pp. 1687–1706. DOI: https://doi.org/10.1155/2019/4368036.

[9] Joshi B., Kumar Sharma A., Singh Yadav N., Tiwari S. (2021). DNN based approach to classify Covid'19 using convolutional neural network and transfer learning. *International Journal of Computers and Applications* [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/1206212X.2021.1983289 (Accessed 02/18/2022).

[10] Zhao Y. (2021). Research on Management Model Based on Deep Learning. *Complexity* [Online]. Available: https://www.hindawi.com/journals/complexity/2021/9997662/ (Accessed 02/18/2022)

R. Yakubovskyi received the Bachelor's degree in Computer Engineering at Lviv Polytechnic National University in 2022.

From 2019 to 2020 he was a domain specialist at Namecheap. Since 2020 and to the present, he has been a Project Manager at various companies, such as ArtCodeVision, CIENCE and Markupus.

Yuriy Morozov received the B. S. degree in Physics at Lviv State University in 1989 and the M.S. degree in Metrology and Radio Engineering at Lviv Polytechnic National University in 1997. His current degree is candidate of technical sciences which he received in 1998.

From 1989 to 1991 he was an engineer at Lviv Research Radio Technical Institute. Since 1991 he started working as a senior engineer at Metrology department of the "Gallar" company. In 1994 he switched to deputy director role at Optimus-Ukraine JV. From 1999 to the present he works as an associate professor of the department of Computer Science.

He is an author of more than 11 articles. His scientific interests cover: creation of systems of complex information protection (CIS), design of virtual communication networks (VPN), means of encoding information, systems of demarcation of access to information, means of analysis of network stability and mechanisms of attack detection.