Concept of Speaker Age Estimation Using Neural Networks to Reduce Child Grooming

Renat Haluska[1,*], Monika Badovska[2], Matus Pleva[1]
[1]*Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Kosice,*
Letna 9, Kosice, Slovak Republic
[2]*Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Kosice,*
Letna 9, Kosice, Slovak Republic
[*]*renat.haluska@tuke.sk; monika.badovska@student.tuke.sk; matus.pleva@tuke.sk*

***Abstract*—This paper focusses on using neural network models to predict the age of social media users based on their voice recordings. The objective is to identify potential risky interactions between minors and adults by comparing the declared and predicted age groups of the users. The paper addresses the selection and training of suitable models and evaluates their effectiveness in age prediction. The results are demonstrated in sample data, where performance metrics are analysed, and possible limitations of the method are identified. Finally, the implications of the results for the safety of minors on social networks are discussed, and suggestions for future research in this area are provided.**

***Index Terms*—Convolutional neural networks; Deep learning; Grooming detection; Human voice; Social networking (online).**

I. INTRODUCTION

The issue of child manipulation requires our attention and joint efforts to understand and address its multi-layered nature. Manipulation, known as grooming, involves the creation of trust with the aim of subsequent exploitation, particularly of vulnerable individuals, such as children [1]. Although grooming is common to associate with sexual exploitation, it also includes a wider spectrum of manipulative behaviour that extends beyond the sexual sphere. This multidisciplinary field combines knowledge from machine learning, voice recognition, and human behaviour analysis. Current technological advances and increased awareness of child protection emphasise the need to prevent and address manipulative behaviour [2].

Grooming takes the form of emotional, social, financial manipulation and sexual exploitation. Recognising these signs is key to protecting children and mitigating potential long-term consequences. A multidimensional view that includes psychological, technological, and sociological aspects is important to comprehensively address the challenges that grooming presents in the context of child abuse [3]. When we speak, our voice carries a wealth of information, including pitch, intonation, speech rate, and phonetic nuances that can be relevant for our age. Although our voices change as we age, capturing these subtle variations is a complex task that requires the performance of advanced neural networks [4].

Deep learning models, particularly recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been at the forefront of this effort. These models can analyse large amounts of spoken data and learn complex relationships between acoustic properties and age [5]. They can identify patterns and nuances that may not be apparent to the human ear. The process begins with the collection of extensive and diverse data sets about the speech of the so-called data sets. These data sets include speakers of different ages, genders, and dialects. The more complex the data, the more robust and accurate the age determination models can be. These data are pre-processed including steps such as extraction of relevant acoustic parameters such as fundamental frequency, resonance, and speech rate.

II. THREATS AND PREVENTION OF GROOMING

The victimisation of young people through sexual abuse has been the focus of study for decades, even before the advent of the Internet [6]. Grooming is now universally understood as a technique to help sexual predators turn their fantasies into reality, online or offline [7].

The term “grooming” was first used in UK legislation in 2003, as part of the Sexual Offences Act [8]. The inclusion of this term was considered progressive, as it allowed the criminalisation of preparatory steps that could lead to sexual abuse of children. However, this law failed in its imprecise definition of sexual grooming. This led to the proposal of the following definition: “The process by which a person prepares a child, they confidants and the environment for the abuse of that child. Specific goals include gaining access to

Manuscript received 18 March, 2024; accepted 8 July, 2024.

This work was partially supported by the Scientific Grant Agency of the Ministry of Education, Research, Development and Youth of the Slovak Republic and the Slovak Academy of Sciences under Grant No. VEGA 2/0165/21; the Cultural and Educational Grant Agency of the Ministry of Education, Research, Development and Youth of the Slovak Republic under Grant No. KEGA 049TUKE-4/2024; partly from the project FakeDetect-Early Fake Profile Detection (Grant No. FBR-PDI-019) financed by Norwegian funds and EEA 2014-2021; partly within the projects supported by the Agency for Research and Development Support under Grants Nos. APVV-22-0261 and APVV-22-0414.

the child, gaining the child's compliance, and maintaining the child's confidentiality to avoid disclosure. This process serves to reinforce the offender's shame schema as it can be used as a way of justifying or denying their actions" [9]. This definition can be applied to the real world or online. The behaviour and goal of grooming remain consistent in all environments, despite possible variations in specific grooming techniques [10].

Children are vulnerable to adult sexual predators because their development of social skills is not yet complete. This makes them less likely to pick up on relevant cues, such as inappropriate remarks that predators use during conversation. Children with low self-esteem, low self-confidence, and greater naivety are at increased risk as they are more likely to be targeted by predators. Sexually curious adolescents, who are more often simply excited, are also more likely to take risks than less curious children, making them a better target group for predators [11].

Child grooming, the deliberate behaviour intended to gain the child's trust and cooperation before engaging in sexual behaviour, is a process that begins with sexual predators choosing a location or a target location that is attractive to children. Later, the grooming process begins, during which predators take a special interest in their child victim to make them feel special with the intention of gaining them trust. As soon as the child begins to trust the predator, predators try to desensitise the child victims to sexual behaviour by introducing a sexual element into the relationship [12].

A series of nine case studies is provided to provide insight into how child grooming occurs and what punishments have imposed on convicted predators. Based on these cases, it appears that the perpetrators use a variety of technologies to facilitate their grooming activities and conduct the grooming activity several months before arranging a physical meeting. Predators have also raised the so-called "fantasy defence", claiming that their actions were expressions of fantasy and did not indicate real intentions [13]. Such defences have been successful in some cases in the United States, creating additional difficulties for prosecutors.

III. Existing Speaker Age Estimation Models

Estimating a person's age from their voice is not a new challenge in the field of speech research. Before this study was conducted, many scientists and researchers have addressed this issue and proposed various models to achieve this goal. In this section, we will briefly review some of these existing approaches.

Kwasny and Hemmerling [14] discuss the use of a speaker embedder network to estimate the age and sex of the speaker. The authors used the Common Voice data set, a large open source data set of multilingual speech recordings. They excluded recordings where the speaker's gender was labelled "other". The training set included data from 54,593 male speakers and 18,099 female speakers. The authors in this paper propose a system for estimating age and gender from speech signals using deep neural networks and transfer learning. Their system achieves a mean absolute error of 5.12 years for men and 5.29 years for women on the TIMIT data set. For gender classification, the system achieves an accuracy of 99.6 % using a d-vector based system.

Ghahremani *et al.* [15] discuss the identification of speakers and language recognition. The authors propose a system that uses an x-vector neural network architecture to estimate the age of a speaker. The system is trained on a data set of speech recordings from NIST SRE08 and evaluated on NIST SRE10. The results show that the proposed system achieves a 12 % improvement in mean absolute error compared to the i-vector baseline.

Stathopoulos, Huber, and Sussman [16] investigate changes in vocal characteristics throughout a person's life. The findings reveal that these alterations are frequently nonlinear and vary depending on sex. Men and women exhibit different nonlinear patterns in fundamental frequency (F0) over their lifetimes. Additionally, the study explores the nonlinear trend in signal-to-noise ratio (SNR) for women, while men demonstrate a linear increase in SNR with age. Notably, the variability in all three acoustic measures is found to be greater in both younger and older speakers.

Asci *et al.* [17] explore the performance of a machine learning model designed to classify age and gender from voice samples. The researchers discovered that the algorithm excelled at differentiating between younger adults (under 25 years of age) and older adults (over 55 years of age). This proficiency was consistent between the two genders. Furthermore, the algorithm effectively distinguished between young men and young women.

A. Deep Learning and Bidirectional Long Short-Term Memory

A. A. Mohammed and Y. F. Al-Irhayim [18], from the University of Tikrit in Iraq, dealt with the estimation of speaker age, which uses the Mel-frequency cepstral coefficients (MFCCs) extraction method for voice analysis, followed by the bidirectional long short-term memory (BiLSTM) method for classification.

The Mel scale is defined as a perceptual measure of frequencies that are the same distance apart. MFCCs are parameters that together form a Mel-frequency cepstrum in which frequency bands are evenly distributed on the Mel scale. MFCCs extract 13 features for each frame. The MFCC response better matches the human auditory system than traditional methods that use linearly distributed frequency bands [19].

The features extracted from the MFCCs are a representation of a relatively stable short-term signal, but due to the temporal variability of speech, a more accurate representation is needed that describes the overall change in the signal over time [20]. These students achieved this by calculating the first derivative (delta) for the extracted characteristics, which considers the difference between the values at two different points in time. Subsequently, the second derivative (delta-delta) was implemented, which is used to represent the signal in a longer time context. Each of these steps produces 13 features for each frame.

Long short-term memory (LSTM) networks are a type of recurrent neural networks (RNN) that can learn from long sequences of data. Unlike traditional RNN networks, LSTM networks can handle long-term dependencies without vanishing gradients [21]. This makes them suitable for tasks such as text analysis, speech recognition, and machine translation. BiLSTM networks are an extension of traditional

LSTM networks, designed specifically to improve performance in series classification tasks [22]. Unlike standard LSTM networks, BiLSTM networks use two independent LSTM networks. One network process the input sequence in its original *order*, while the other processes its reversed copy. This approach allows BiLSTM networks to capture contextual information from both directions of the sequence, resulting in faster and more comprehensive learning compared to standard LSTM networks.

The results showed that the best accuracy was achieved when the audio recording was 10 seconds long and the batch size of the audio features for each training epoch was 50 attributes. The accuracy rate for the age estimation network reached 94.008 %.

B. Comparison of Human and Machine Speaker Age Estimation

Fascinated by the accuracy of age estimation, Huckvale and Webb [23] from University College London embarked on a research project that compared the accuracy of age estimation of speakers by humans and machine learning.

A web-based data collection protocol was used to obtain human age estimation data. Listeners could listen to each test recording and then estimate their age using a sliding scale from 15 to 80 years. Estimates were recorded as integer years. The recordings were presented in a random order that was different for each listener. Listeners could estimate their age at any time while the recording was playing, or they could listen to the audio multiple times.

It was investigated whether the gender and age of the speaker and listener affect the accuracy of the age estimation. The gender of the speaker was found to have no significant effect. However, the age group of a speaker had a significant effect. The age estimates of the speakers of the 20–29 age group were significantly better than the estimates of other groups. The gender of the listener had no significant effect. The age group of the listeners had an effect, with people in the 40–49 and 60–69 age groups achieving worse estimates as a group of 20–29 years.

In the study, they compared human listeners and machine learning in estimating the age of a speaker. They showed that humans and machines achieve similar results, with machines having a slight advantage (smallest machine error: 8.64 years, human error: 9.79 years). Both groups had problems with age estimation at the edges of the age limit. Increasing the number of older people in training did not help. Larger data sets with diverse age groups are likely to yield more accurate estimates in the future.

IV. DATA PREPROCESSING

A source from the Kaggle website, which focusses on data sets for machine learning, was used to select the data set. In addition to forums about various techniques and tools, there are also freely available data sets, from which we chose "Common Voice" [24]. It contains 9134 audio recordings of Slovak speakers and a *cvc* file with their paths and the age of the person recorded in decades whose voice is on the track, as well as information such as gender and accent with over 31 hours of labelled audio data [25],

The data itself are raw, and certain actions need to be performed on the data to properly handle it when training the model. The performance of such actions is called "data preprocessing" and is necessary in training if we want to achieve full-fledged results.

The first step in preprocessing is to view the data and understand the connections between them. In this case, a folder with audio tracks and a table file were available, where the first column refers to the path to the given audio track, and the other columns contain information about the given recording. Within this issue, we will only be interested in information about the age of the given person and, therefore, apart from age and the path to the recordings, the remaining columns are not important. When working with data sets that are downloaded, you must consider the incompleteness of the data, which in our case are not specifying the age in the table file, or an invalid path to the recording or a path that does not exist.

Label encoding is the formatting of data that are in textual form into numerical form, where a numerical value is assigned to each unique occurrence of the label. It is an unavoidable action for the proper functioning of the programme. Table I shows the age encoding from text to number.

TABLE I. AGE CATEGORIES BEFORE AND AFTER ENCODING.

Text	Number
Teens	0
Twenties	1
Thirties	2
Forties	3
Fifties	4
Sixties	5
Seventies	6
Eighties	7

For the training process, it is necessary to first convert the audio data into a format that is understandable by neural networks. Using the *Librosa* library [26], it is possible to extract the properties of audio tracks and create spectrograms from them, which are a visual representation of frequencies and thus of sound. Different audio tracks can produce spectrograms of different sizes. To achieve consistent data, it is necessary to ensure that all spectrograms have a uniform size, e.g., 128 by 128 pixels [27]. In Fig. 1, you can see a spectrogram directly from the data, specifically from the tenth audio recording.

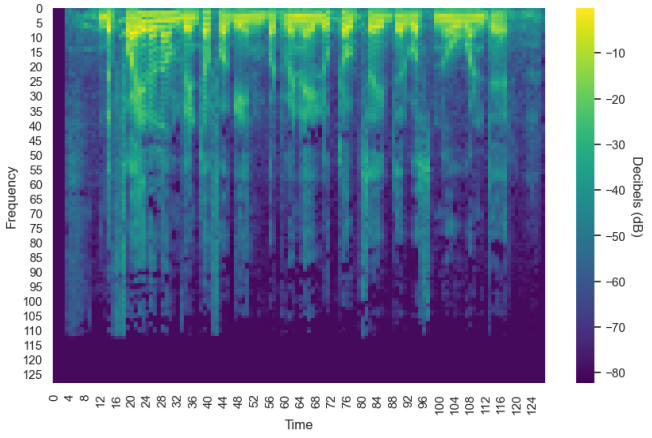


Fig. 1. Spectrogram of a sample from the data set: Features of the tenth sample.

Before training the model, it is important to divide it into a training set and a test set. Splitting the data allows you to verify the performance of the model in new, unseen data, which provides an estimate of its ability to generalise to the data. Testing a set that consists of data that the model did not see during training serves as an estimate of the model's performance in real situations and helps assess how well it could perform in practice. Without a test set, there is a risk that the model will learn from the data itself instead of generalising, which is also called "overtraining".

After the data have been split, it is essential to check the shape and distribution of each part of the data set. Understanding the size and composition of the training, validation, and test set provides valuable information about the characteristics of the data set and helps ensure that each part adequately represents the overall data set. In addition, it makes it possible to identify potential problems, such as imbalances between classes or data bias.

V. MODELS OF NEURAL NETWORKS

When choosing neural network models, deep neural networks (DNN) and convolutional neural networks (CNN) were chosen. These models are popular in the world of deep learning and are widely used in various fields such as image recognition, natural language processing, or time series prediction.

The initial configuration was defined for CNN architecture with two convolutional layers, each used 3×3 kernels and rectified linear unit (ReLU) activation. The number of filters used (32 and 64) controlled the complexity of learnt features. Experimentation with different filter counts (16, 48, 128) was beneficial depending on the size of the data set and the desired level of detail in the extracted features. Similarly, exploring various kernel sizes (1×1, 5×5) was useful to capture features of different scales. The final size of the dense layer (128) determined the model's capacity to learn nonlinear relationships between features. Adjusting this value (e.g., 64, 256) impacted performance and required careful consideration to avoid overfitting with limited data.

The provided analysis explores potential adjustments for the parameters within an image classification CNN model. Although the current configuration uses Adam optimiser, ReLU activation, and caters to 128×128 greyscale images with eight classes, there is room for optimisation. Exploring alternative optimisers such as root mean squared propagation (RMSprop) or stochastic gradient descent (SGD) with learning rate scheduling, or using Leaky ReLU for specific tasks, could improve performance. Remember to adjust the input shape and the number of classes to match your data set. Additionally, experimenting with batch size, epochs, and regularisation techniques can help fine-tune the model and prevent overfitting, especially for limited data.

Both models were trained using the "Adam" optimisation algorithm, which is commonly used for fast neural network training. These models were trained in the training set using the "sparse categorical cross-entropy" loss function and evaluated using accuracy and loss.

To achieve better results and improve the performance of the models, the next stage of the process involves tuning the hyperparameters. This procedure makes it possible to optimise model settings and obtain better results, which can lead to more accurate predictions and more powerful models. Hyperparameter tuning is the process of optimising model performance by experimenting with different combinations of values of the so-called "hyperparameters". These hyperparameters are not learnable directly from the data, but affect the way the model learns.

By using the GridSearch function, which makes it possible to systematically examine various combinations of model parameters, various options for optimising the model are presented. In this case, these are different optimisation algorithms, activation functions, number of epochs, and size of the training group. This approach made it possible to systematically explore a wide range of possible model configurations and find the best combinations of parameters that lead to the best results.

Most systems and Web applications store audio files in data bases, so simply loading the file should be sufficient. However, before the prediction itself, it is important to edit this file to a spectrogram with the same dimensions and sound properties as those used in training the model. Subsequently, the model can be used to predict the age of a person from an audio recording. This procedure is a key step in analysis and modelling and allows for effective use of learnt patterns on new data.

After the initial training of the DNN and CNN models, we got the first results. These results, presented in Table II, provide an overview of the accuracy and loss values achieved during training.

TABLE II. RESULTS OF CNN AND DNN MODELS.

Model	Accuracy	Loss
CNN	0.52	4.15
DNN	0.41	1.67

Subsequently, an analysis of the results was performed using the precision, recall, and F1-score metrics, (see Table III); thus, we obtained a more detailed view of the performance of the models for individual classes using the classification report, as shown in Figs. 2 and 3.

TABLE III. CNN AND DNN METRICS.

Metrics	Precision	Recall	F1-Score
CNN	0.5240	0.5218	0.5185
DNN	0.4337	0.3669	0.3604

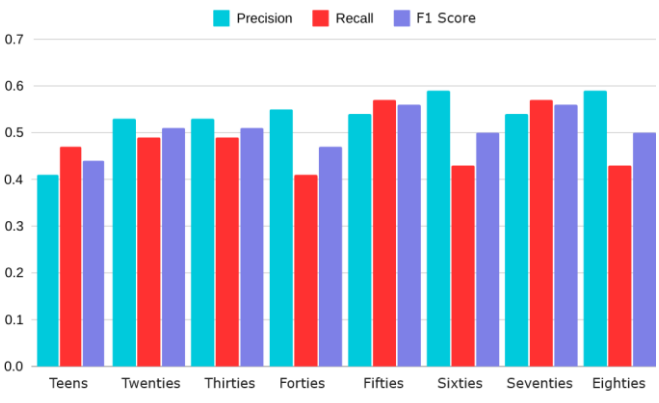
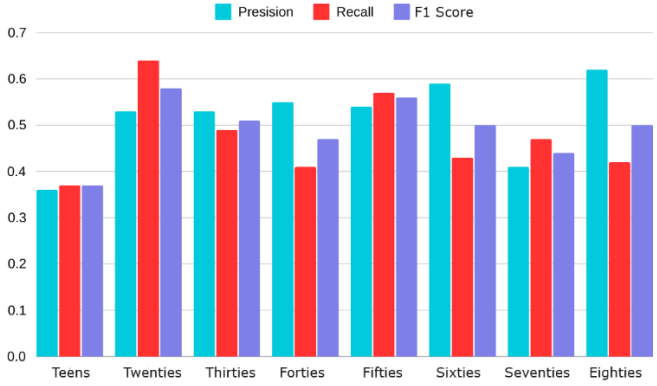


Fig. 2. Graphic representation of CNN classification report.

Fig. 3. Graphic representation of DNN classification report. The goal of tuning was to find the optimal parameter

settings for each model (DNN and CNN models) to achieve maximum classification accuracy.



The tuning results are summarised in Fig. 4 which shows the different parameters that were tested and their effect on the classification accuracy for each model.

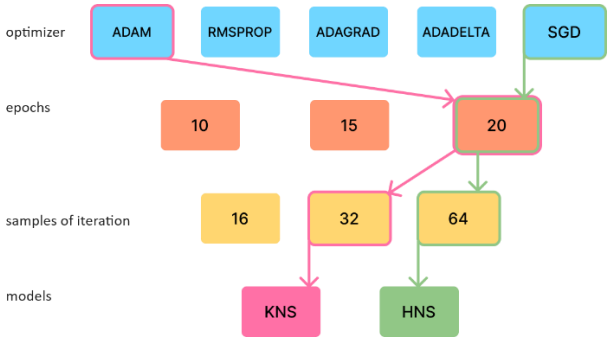


Fig. 4. Hyperparameter tuning diagram.

In Fig. 4, we can see:

– Optimiser (blue): Algorithm used to optimise model parameters during training. Tried by: {Adam, RMSProp, AdaGrad, Adadelta, SGD};

– Epochs (orange): Number of passes through the entire training data set during training. Tried: {10, 15, 20};

– Samples per iteration (batch size) (yellow): Number of training samples processed in one iteration. Tested: {16, 32, 64}.

The prediction procedure consists of the following steps:

– Saving models;

– Load models;

– Data preprocessing;

– Age prediction.

It is clear from Fig. 4 that the optimal parameter settings are different for different models. For example, the DNN model achieved the best accuracy with the “Adam” optimiser at twenty epochs with a sample size of 32. In contrast, the DNN model achieved the best accuracy with the “SGD” optimiser also at twenty epochs, but with a sample size of 64.

The results of age predictions on a random selection of samples from each class are presented in the form of a confusion matrix. This matrix provides an overview of the quality of the predictions. On the diagonal of the matrix, there are the correctly classified samples, while off the diagonal, there are the incorrectly classified samples. Using this matrix, it is possible to identify which age classes were often confused and it gives us an insight into the error rate of the models.

Predication of majority classes is more accurate than for minority classes as presented in Fig. 5. DNN model in Fig. 6 decently predicted the minority classes, and out of six samples for the majority class, namely the class of twenty-somethings, he did not correctly predict even one, although he estimated the classes close.

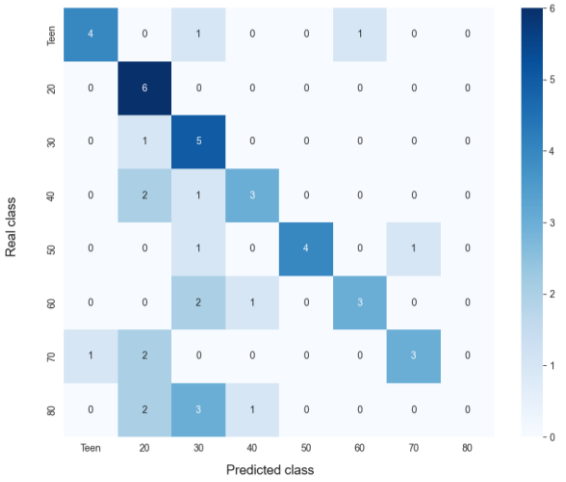


Fig. 5. Confusion matrix for CNN model.

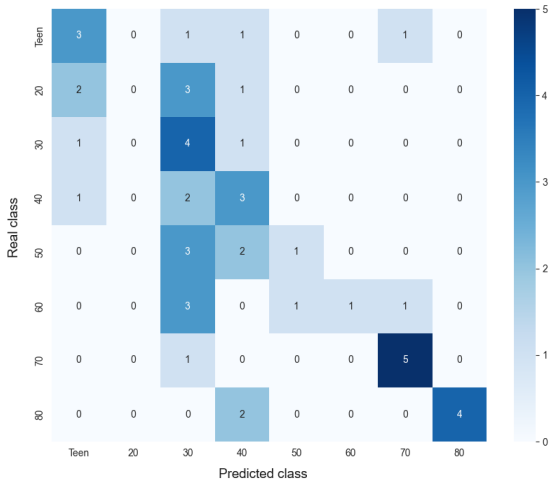


Fig. 6. Confusion matrix for DNN model.

In general, the CNN model predicts better and is more suitable for the grooming issue, since the majority classes are the age ranges of common users of social networks and the Internet in general.

VI. PROPOSAL FOR IMPLEMENTATION

The trained model could also be used in practice as part of prevention against grooming on children. There are several ways of implementation, e.g., age verification by voice already during user registration, a way to verify users who will not have limited options, or just as a simple note under the profile of a user whose estimated age does not match the age entered. This information can be useful not only for victims themselves, if they do not know the real age of the person with whom they are texting, but also for the authorities, who would be informed if a child’s conversation with someone much older is detected.

The possible functioning of the system is shown in Fig. 7,

and the age detection in a private conversation between two users is shown in Fig. 8, where possible functionality is shown within a private conversation between two users. The moment the application has access to the audio recording of the user's voice, it predicts the age range using the model and compares it with the age on the profile. If the age at registration matches the recognised age, the conversation continues, but if this is not the case, a warning will be displayed that will alert the user to this fact and give him the option to block the user or continue the conversation.

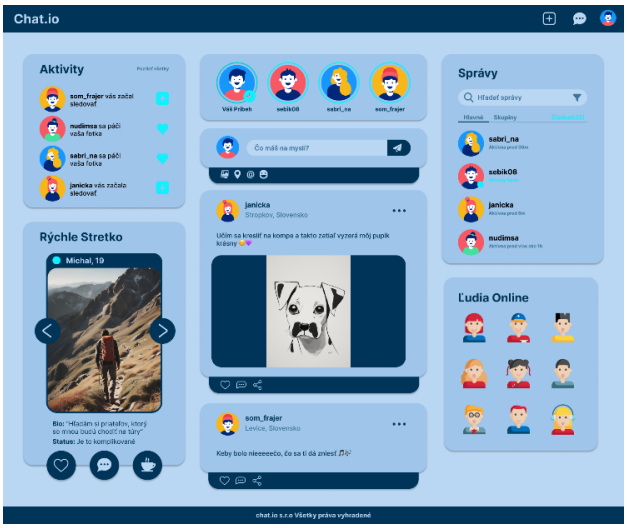


Fig. 7. Functions of the proposed system (in Slovak).

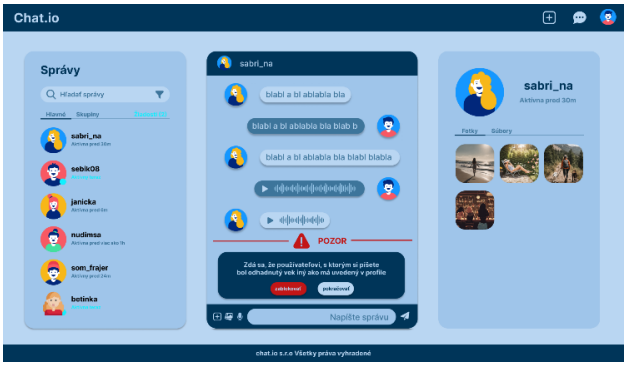


Fig. 8. Example of detection (“Alert” - “Pozor” in Slovak).

This proposal is nonintrusive and gives users the choice of whether they want to communicate with someone who is a different age than they indicate, but as part of the prevention against grooming, notification of victims would also mean immediate contact of the authorities.

The design of this system implementation does not take into account a user who does not share his voice track. A possible solution would be mandatory proof of voice, e.g., during registration. It could also be optional, but in that case the user could have limited access and not be able to, e.g., send photos and text with users who have not friended them first. Such an implementation would be more difficult to implement on social networks that already exist and do not have user verification implemented.

VII. Discussion

This paper investigated the use of neural networks to estimate the age of a speaker from the voice to detect child grooming by proposing two models: a convolutional neural network (CNN) and a deep neural network (DNN). While the CNN model achieved a higher overall accuracy (0.57) in age prediction, the DNN achieved an accuracy of 0.44. It was interesting to find that DNN was better able to estimate the age categories less represented (minority classes). The analysis using precision, recall, and F1-score metrics gave us a more detailed view of the performance of the models for individual age groups. Despite the implementation of methods to ensure an even distribution of data during training, the imbalance of the data set still seems to play a role during prediction. A solution would be to train on a more balanced data set or use a combination of downsampling with oversampling as opposed to just oversampling.

Another approach which exists is that moderated applications dedicated to children as a safe place to interact are monitored by the company moderators. The company is declaring that the interaction inside the application provides a safe place for children. The company develops solutions to monitor users, chats, and voice chats to detect signs of child grooming. The application provides alarms to the moderator about possible adults in text chats [28], [29]. Similar alarms could be generated by the proposed system. In this case, only two groups will be evaluated: below 20 and more than 20. In this case, the CNN model provides 93,87 % and if we focus only on adults detected as under 20, the model provides 97,67 % accuracy, which could be a great starting point for the demo application. Alarms could be provided to the user or to the moderator for double check.

VIII. Conclusions

Based on the age group estimated by the models, a simple grooming detection system was designed. This system can warn of potentially risky situations where one participant in the conversation is classified as a teenager and the other participant is in the age group no older than twenties. The demo system for detecting adults in children audio chats based on CNN provides 97,67 % accuracy. Similar adult/child voice detection system was to our knowledge not described yet, so it provides a starting point for the new area of research. Similarly, we are working on synthetised/morphed voice detection to prevent to use nonreal voice in the voice chats.

The work used a relatively large training set of data, this set was unbalanced in the representation of age categories. Despite this drawback, the results showed the promising potential of neural networks for estimating a speaker’s age from his voice. In the future, exploring other audio processing techniques, such as data augmentation and normalisation, could improve the performance of the models and mitigate the impact of data imbalance. The detection system could be further improved by analysing the text of the conversation in addition to age estimation. The proposed models and detection system represent a promising step forward to reduce this serious online danger, especially in today’s world where children often gain access to the Internet at an early age.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

REFERENCES

[1] C. H. Ngejane, G. Mabuza-Hocquet, J. H. P. Eloff, and S. Lefophane, "Mitigating online sexual grooming cybercrime on social media using machine learning: A desktop survey", in *Proc. of 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, 2018, pp. 1–6. DOI: 10.1109/ICABCD.2018.8465413.

[2] J. H. Alves *et al.*, "Detecting relevant information in high-volume chat logs: Keyphrase extraction for grooming and drug dealing forensic analysis", in *Proc. of 2023 International Conference on Machine Learning and Applications (ICMLA)*, 2023, pp. 1979–1985. DOI: 10.1109/ICMLA58977.2023.00299.

[3] M. Rybnicek, R. Poisel, and S. Tjoa, "Facebook Watchdog: A research agenda for detecting online grooming and bullying activities", in *Proc. of 2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2013, pp. 2854–2859. DOI: 10.1109/SMC.2013.487.

[4] M. Fairhurst, M. Erbilek, and M. Da Costa-Abreu, "Selective review and analysis of aging effects in biometric system implementation", *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 3, pp. 294–303, 2015. DOI: 10.1109/THMS.2014.2376874.

[5] H. Rong, "Exploration of lightweight neural network architectures for sentiment analysis", in *Proc. of 2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 2023, pp. 312–316. DOI: 10.1109/ICICML60161.2023.10424736.

[6] G. M and A. Vidhya, "Optimized hybrid model using machine learning to combat the prevalence of cybercrime", in *Proc. of 2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2023, pp. 739–746. DOI: 10.1109/ICECA58529.2023.10395218.

[7] M. A. Fauzi and P. Bours, "Ensemble method for sexual predators identification in online chats", in *Proc. of 2020 8th International Workshop on Biometrics and Forensics (IWBF)*, 2020, pp. 1–6. DOI: 10.1109/IWBF49977.2020.9107945.

[8] "Sexual Offences Act 2003", legislation.gov.uk, 2012. [Online]. Available: https://www.legislation.gov.uk/ukpga/2003/42/section/15

[9] S. Craven, S. Brown, and E. Gilchrist, "Sexual grooming of children: Review of literature and theoretical considerations", *Journal of Sexual Aggression*, vol. 12, no. 3, pp. 287–299, 2006. DOI: 10.1080/13552600601069414.

[10] H. Whittle, C. Hamilton-Giachritsis, A. Beech, and G. Collings, "A review of online grooming: Characteristics and concerns", *Aggression and Violent Behavior*", vol. 18, no. 1, pp. 62–70, 2013. DOI: 10.1016/j.avb.2012.09.003.

[11] Y.-Q. Wen, Y.-W. Chen, and Y.-T. Zhang, "Research on the structural relationship between parent-child relationship and the usage of smartphone, academic procrastination and self-esteem based on structural equation model under the background of big data", in *Proc. of 2022 3rd International Conference on Education, Knowledge and Information Management (ICEKIM)*, 2022, pp. 476–481. DOI: 10.1109/ICEKIM55072.2022.00113.

[12] J. Jurinić and T. Ramljak, "Sexual exploitation or child pornography: Terminological analysis in Criminal Codes of Southeast European countries", in *Proc. of 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2021, pp. 1502–1510. DOI: 10.23919/MIPRO52101.2021.9596657.

[13] R. Smith, P. Grabosky, and G. Urbas, "Cyber criminals on trial", *Criminal Justice Matters*, vol. 58, no. 1, pp. 22–23, 2008. DOI: 10.1080/09627250408553240.

[14] D. Kwasny and D. Hemmerling, "Gender and age estimation methods based on speech using Deep Neural Networks", *Sensors*, vol. 21, no. 14, p. 4785, 2021. DOI: 10.3390/s21144785.

[15] P. Ghahremani *et al.*, "End-to-end deep neural network age estimation", in *Proc. of Interspeech 2018*, 2018, pp. 277–281. DOI: 10.21437/Interspeech.2018-2015.

[16] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age", *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1011–1021, 2011. DOI: 10.1044/1092-4388(2010/10-0036).

[17] F. Asci *et al.*, "Machine-learning analysis of voice samples recorded through smartphones: The combined effect of ageing and gender", *Sensors*, vol. 20, no. 18, p. 5022, 2020. DOI: 10.3390/s20185022.

[18] A. A. Mohammed and Y. F. Al-Irhayim, "Speaker age and gender estimation based on deep learning bidirectional Long-Short Term Memory (BiLSTM)", *Tikrit Journal of Pure Science*, vol. 26, no. 4, pp. 67–84, 2021. DOI: 10.25130/tjps.v26i4.166.

[19] Z. Hong, "Speaker gender recognition system", M.S. thesis, Department of Communications Engineering, University of Oulu, Oulu, Finland, 2017.

[20] P. Nantasri *et al.*, "A light-weight artificial neural network for speech emotion recognition using average values of MFCCs and their derivatives", in *Proc. of 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2020, pp. 41–44. DOI: 10.1109/ECTI-CON49241.2020.9158221.

[21] M. Bkassiny, "A deep learning-based signal classification approach for spectrum sensing using long short-term memory (LSTM) networks", in *Proc. of 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2022, pp. 667–672. DOI: 10.1109/ICITISEE57756.2022.10057728.

[22] M. Ibrahim, K. M. Badran, and A. E. Hussien, "Artificial intelligence-based approach for univariate time-series anomaly detection using hybrid CNN-BiLSTM model", in *Proc. of 2022 13th International Conference on Electrical Engineering (ICEENG)*, 2022, pp. 129–133. DOI: 10.1109/ICEENG49683.2022.9781894.

[23] M. Huckvale and A. Webb, "A comparison of human and machine estimation of speaker age", in *Statistical Language and Speech Processing. SLSP 2015. Lecture Notes in Computer Science()*, vol. 9449. Springer, Cham, 2015, pp. 111–122. DOI: 10.1007/978-3-319-25789-1_11.

[24] "Common Voice", kaggle.com, 2017. [Online]. Available: https://www.kaggle.com/datasets/mozillaorg/common-voice/data

[25] "Common Voice by Mozilla", commonvoice.mozilla.org. [Online]. Available: https://commonvoice.mozilla.org/sk/languages

[26] B. McFee *et al.*, "Audio and music signal analysis in Python", in *Proc. of the 14th Python in Science Conf. (SciPy 2015)*, 2015, pp. 1–7. DOI: 10.25080/Majora-7b98e3ed-003.

[27] M. F. Khan, R. N. S. Kumar, T. Patil, A. Reddy, V. Mane, and S. Santhoshkumar, "Neural network optimized medical image classification with a deep comparison", in *Proc. of 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 2022, pp. 11–15. DOI: 10.1109/ICAISS55157.2022.10011109.

[28] M. A. Fauzi, S. Wolthusen, B. Yang, P. Bours, and P. Yeng, "Identifying sexual predators in chats using SVM and feature ensemble", in *Proc. of 2023 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, 2023, pp. 1–6. DOI: 10.1109/ETNCC59188.2023.10284950.

[29] P. R. Borj, K. Raja, and P. Bours, "Detecting online grooming by simple contrastive chat embeddings", in *Proc. of the 9th ACM International Workshop on Security and Privacy Analytics (IWSPA'23)*, 2023, pp. 57–65. DOI: 10.1145/3579987.3586564.