https://doi.org/10.31891/csit-2026-1-15

Oleh TOMASHEVSKYY
Candidate of Technical Sciences,
Associate Professor of Artificial Intelligence Systems Department, Lviv Polytechnic National University
https://orcid.org/0009-0009-9134-7520
e-mail: Oleh.M.Tomashevskyi@lpnu.ua

Orest TKACHUK
Phd Student of Artificial Intelligence Systems Department, Lviv Polytechnic National University
https://orcid.org/0009-0007-1216-6787
e-mail: orest1040@gmail.com

Received: 22/01/2026
Accepted:28/02/2026
Published:26/03/2026

UDC 004.8:004.93

CONVOLUTIONAL NEURAL NETWORK-BASED SOUND SOURCE SEPARATION IN THE TIME-FREQUENCY DOMAIN

This paper addresses the problem of sound source separation in mixed audio signals in the time-frequency domain. The study considers the application of convolutional neural networks for isolating individual acoustic components from complex audio mixtures where multiple sources overlap in both time and frequency. The presence of such overlap significantly complicates the separation process and increases the requirements for stability and structural consistency of the applied models. The proposed approach is based on transforming audio signals using the Short-Time Fourier Transform and representing audio mixtures as spectrograms that preserve both temporal and spectral characteristics of sound components. A binary masking strategy is applied to the resulting representations to structurally simplify the separation task. A convolutional neural network is employed to predict masks corresponding to individual sound sources such as vocals, bass, drums, and other components. This masking formulation enables selective extraction of spectral regions associated with specific sources and supports the implementation of a hybrid processing scheme that combines elements of classification and regression within a unified neural architecture. The research methodology includes the design of the network architecture, preparation of spectrogram-based input data, model training on multi-source audio mixtures, and validation of separation quality using reconstruction consistency criteria. Particular attention is paid to ensuring stable convergence of the model and preserving meaningful acoustic patterns within the predicted masks. The findings demonstrate stable isolation of sound components and consistent performance across training and validation datasets. Quantitative evaluation shows separation accuracy of 0.772 for vocals, 0.766 for drums, 0.944 for bass, and 0.764 for other sources, with corresponding mean squared error values ranging from 0.044 to 0.203 across evaluated categories. The highest performance was achieved for bass isolation due to the distinct low-frequency spectral structure of this source. Signal-level evaluation using SI-SDR, SDR, and SNR metrics produced values ranging from -1.24 to 4.10 dB (SI-SDR), -0.26 to 4.59 dB (SDR), and 1.16 to 5.09 dB (SNR), with the highest values observed for bass and vocal sources, consistent with the accuracy-based results. The results confirm the effectiveness of integrating binary masking with convolutional processing of spectrograms for computationally efficient sound source separation. The proposed approach, implemented using a compact neural architecture with 323,233 trainable parameters, can be applied in music production systems, speech enhancement solutions, intelligent audio analysis platforms, and other audio processing environments requiring reliable and lightweight separation mechanisms.

Keywords: computer science, artificial intelligence, convolutional neural networks, audio data analysis, audio signal processing, sound source separation.

Introduction

The rapid development of digital technologies and multimedia systems has significantly increased the volume of audio data used in various fields, including music production, speech processing, and intelligent information systems. In many practical scenarios, audio recordings contain multiple overlapping sound sources, making the task of isolating individual components technically complex and computationally demanding.

Sound source separation aims to extract separate acoustic components from a mixed audio signal. In polyphonic compositions, different instruments and vocal parts are combined within a single waveform, and their spectral components overlap in the time-frequency domain. This overlap complicates direct signal reconstruction and requires the development of efficient and stable processing approaches.

Modern audio processing systems increasingly rely on machine learning techniques for solving complex signal analysis

tasks. Neural network architectures enable automated extraction of relevant features from data representations and allow the construction of adaptive models capable of handling high-dimensional inputs. Time-frequency representations of audio signals provide structured information about both temporal and spectral characteristics, which makes them suitable for machine learning-based processing.

Despite significant progress in this area, the problem of designing computationally efficient and structurally consistent separation systems remains relevant. The development of approaches that simplify spectral reconstruction while preserving meaningful acoustic patterns is, therefore, an important research direction.

The purpose of this paper is to develop a machine learning approach to sound source separation in the time-frequency domain that provides a stable and practical way to isolate individual sound components from complex audio mixtures. The proposed method combines spectrogram-based signal representation with convolutional neural network processing and formulates separation as binary mask prediction instead of direct spectral reconstruction. By incorporating localized time-frequency context and a controlled training strategy, the approach ensures consistent mask estimation and reliable reconstruction of target sources.

The main scientific contribution of this work is the investigation of the applicability of convolutional neural networks to the problem of sound source separation in complex audio mixtures where multiple sources overlap in the time-frequency domain. In particular, the study introduces a formulation of the separation task based on a hybrid regression-classification approach, where convolutional neural networks operate on spectrogram representations and estimate binary masks indicating the presence or absence of target sources in individual time-frequency bins. This formulation simplifies the learning problem by shifting the focus from precise signal reconstruction to the identification of dominant spectral components associated with individual sources. As a result, the proposed representation enables the model to exploit localized spectral structures, improves the stability of the learning process, and allows the use of relatively compact convolutional architectures for the structured separation of overlapping sound sources.

The paper is structured as follows. Section 2 presents related works and formulates the problem statement. Section 3 defines the research aim and objectives. Section 4 describes the proposed methodology. Section 5 presents the research results. Section 6 discusses the obtained results. The last section concludes the paper.

Related works

In order to develop a sound source separation system, it is necessary to analyze existing approaches to processing mixed audio signals. Sound source separation has been studied within signal processing and machine learning, and a number of approaches can be identified based on different theoretical and computational principles.

Early research in audio source separation is closely related to blind source separation and statistical signal analysis. These approaches aim to recover individual signal components from mixtures without access to isolated reference signals. Methods based on Bayesian risk minimization and statistical modeling are presented in [3]. Harmonic partial reconstruction and separation of overlapping components are discussed in [6]. Block-wise modeling and extraction of independent vector components from underdetermined mixtures are investigated in [18]. Such approaches rely on assumptions about independence, sparsity, or structural properties of signals. While they provide an important theoretical basis, their performance may decrease in polyphonic music where spectral components strongly overlap.

A significant group of methods is based on matrix factorization and sparse representations in the time-frequency domain. In these approaches, a spectrogram representation of the signal is decomposed into spectral bases and activation coefficients. Active-set algorithms for non-negative representations are proposed in [10]. The use of non-negative matrix factorization with spectral masks is examined in [15]. Minimum-volume beta-divergence NMF for blind audio separation is presented in [19], and optimized complex sparse tensor deconvolution methods are described in [12]. Although these techniques can effectively model spectral structures, their quality depends on the choice of divergence measures, constraints, and initialization procedures. When several instruments share similar frequency ranges, matrix factorization may not provide clear separation of individual sources.

Independent component analysis combined with time-frequency decomposition represents another related direction. An ICA-based time-frequency approach for single-channel source separation is presented in [8]. Such methods attempt to separate components by exploiting statistical independence in the transformed domain. However, their effectiveness depends on the validity of independence assumptions and on the characteristics of the analyzed signals.

With the development of deep learning, neural network-based approaches have become increasingly common in source separation tasks. Deep neural networks allow learning acoustic patterns directly from data and reduce reliance on strict analytical assumptions. A regression-based approach for single-channel speech separation is described in [4], and a two-stage neural architecture is proposed in [16]. Improvement of mask-based separation using deep neural networks is investigated in [23], while multi-task learning strategies for audio source separation are presented in [24]. These models demonstrate improved performance compared to classical statistical and matrix factorization techniques, especially when trained on large annotated datasets.

Convolutional neural networks are particularly suitable for processing spectrogram representations due to their ability to capture local spectral patterns and hierarchical features. Spectrogram-based neural models form the

basis of many modern separation systems[4; 16; 23; 24]. In addition to frequency-domain approaches, end-to-end time-domain architectures have also been proposed. Conv-TasNet [20] demonstrates high separation quality in speech processing tasks and illustrates the effectiveness of direct waveform modeling without explicit time-frequency transformation. However, time-frequency representations remain widely used in source separation due to interpretability and ability to expose localized spectral structures, which can be effectively processed using relatively simple convolutional architectures.

Audio source separation remains an actively studied problem, and recent research continues to explore neural approaches for different types of audio signals. Several recent studies focus on speech and singing voice separation using deep neural architectures designed to isolate a single dominant source from an audio mixture [2; 13]. Other works investigate the separation of specific instrument groups, such as individual percussion components within drum recordings [7]. However, many of these approaches focus on separating one specific type of source. The separation of polyphonic musical mixtures containing multiple overlapping sources remains a more challenging problem.

In recent years, practical implementations of source separation have been integrated into software tools and commercial products. The Spleeter library [17] provides pre-trained neural models for music source separation and is widely used in research and practical applications. Commercial software solutions such as iZotope [22], SpectralLayers [9], Acon Digital [1], VirtualDJ [11], and Algoriddim [14] incorporate separation functionality into audio production workflows. The availability of these tools confirms the practical relevance of separation methods and demonstrates demand for reliable and computationally efficient solutions.

Despite the diversity of existing approaches, several challenges remain. First, polyphonic audio mixtures contain sources with overlapping spectral components, which complicates accurate isolation. Second, spectrogram representations are high-dimensional and may increase computational complexity during training and inference. Third, direct reconstruction of spectral magnitudes may be sensitive to training instability and can introduce artifacts in reconstructed signals. Therefore, the development of machine learning approaches that ensure stable and structurally consistent separation while retaining computational feasibility remains relevant.

Thus, the problem of developing a machine learning approach to sound source separation in the time-frequency domain remains relevant. It is necessary to design a method that ensures stable isolation of multiple acoustic components in polyphonic mixtures, lowers computational complexity associated with high-dimensional spectrogram representations, and retains structural consistency of reconstructed signals. The solution to this problem should provide reliable separation performance and be suitable for practical implementation in modern audio processing systems.

Research aim and objectives

The analysis of existing approaches to sound source separation demonstrates that, despite the diversity of statistical and neural methods, the problem of reliably and computationally efficiently isolating acoustic components in polyphonic mixtures remains relevant. Modern separation systems may provide high-quality results under specific conditions; however, challenges related to spectral overlap, computational complexity, and stability of reconstruction continue to limit practical application in certain scenarios.

The aim of this study is to develop a machine learning approach to sound source separation in the time-frequency domain that ensures stable isolation of multiple acoustic components and retains the structural consistency of reconstructed signals.

The proposed approach is based on the use of spectrogram representations of mixed audio signals and convolutional neural network processing. Particular attention is given to the formulation of a binary masking strategy that simplifies the separation task while preserving meaningful acoustic structures in the reconstructed components. The method is designed to operate within a supervised learning framework using a prepared dataset of polyphonic audio fragments.

To achieve the stated aim, the following objectives are defined:

- analyze and prepare a dataset of mixed audio signals suitable for supervised training;
- construct a time-frequency representation that preserves both spectral and temporal characteristics of the signal;
- formulate a binary masking strategy for separating individual sound sources in the spectrogram domain;
- design a convolutional neural network architecture for predicting separation masks based on local time-frequency patterns;
- implement the training procedure and evaluate the performance of the developed model using appropriate quantitative metrics.

The achievement of these objectives provides the methodological foundation for the development and evaluation of the proposed sound source separation system.

Materials and methods

This section describes the materials and methods used for the development of the proposed sound source separation system. The methodology includes preparation of the audio dataset, construction of time-frequency

representations, formulation of separation masks, design of a convolutional neural network architecture, and implementation of the training procedure.

The study is based on the MUSDB18 [21] dataset (Music Source Separation Database 2018), which is specifically designed for research in music source separation. This dataset provides multitrack musical recordings with isolated source components, enabling supervised training and evaluation of separation algorithms.

MUSDB18 contains 150 musical compositions of various genres and durations. Each composition consists of a mixture track and five isolated source tracks corresponding to different musical components, such as vocals, drums, bass, and other instruments.

An example of a multitrack structure from the MUSDB18 dataset is illustrated in Fig. 1. This structure allows direct comparison between the mixed signal and its individual sources. The dataset is divided into training and test subsets, comprising 100 and 50 compositions, respectively. Such a split ensures sufficient material for model development while preserving independent data for evaluation.

File_Path	File_Name	Music_Length	Sample_Rate	MIXTURE	DRUMS	BASS	OTHER	VOCALS
wav\train\...	A Classic Education - NightOwl.stem.wav	171.247166	22050	[0.01864624, 0.06594849, 0.111816406, 0.142761...	[-0.00881958, 0.011566162, 0.018341064, 0.0016...	[0.11306763, 0.11557007, 0.11785889, 0.1188354...	[-0.011749268, -0.014129639, -0.015472412, -0....	[-0.07418823, -0.046295166, -0.007659912, 0.03...
wav\train\...	Actions - Devil's Words.stem.wav	196.626576	22050	[-0.13696289, -0.10281372, -0.0680542, -0.0354...	[0.0014038086, -3.0517578e-05, -0.0025634766, ...	[-0.027282715, -0.026947021, -0.025939941, -0....	[0.0107421875, 0.016204834, 0.014953613, 0.007...	[-0.124420166, -0.09283447, -0.05368042, -0.01...
wav\train\...	Actions - One Minute Smile.stem.wav	163.375601	22050	[-0.032440186, -0.018157959, 0.007659912, 0.02...	[-0.099731445, -0.08331299, -0.07211304, -0.09...	[0.043060303, 0.04269409, 0.042266846, 0.04187...	[0.05505371, 0.013977051, 0.015838623, 0.05453...	[-0.027282715, 0.009490967, 0.023468018, 0.014...
wav\train\...	Actions - South Of The Water.stem.wav	176.610975	22050	[-0.09185791, -0.07498169, -0.06341553, -0.058...	[-0.012634277, -0.003326416, -0.0015563965, -0...	[-0.053344727, -0.05380249, -0.054718018, -0.0...	[-0.038146973, -0.0289917, -0.01953125, -0.010...	[0.011962891, 0.012054443, 0.012298584, 0.0219...
wav\train\...	Aimee Norwich - Child.stem.wav	189.080091	22050	[0.039367676, -0.032318115, -0.059265137, -0.0...	[-0.0289917, -0.030090332, -0.03253174, -0.028...	[-0.006866455, -0.0033569336, 0.00015258789, 0...	[0.02218628, 0.02166748, 0.012176514, -0.00848...	[0.05606079, -0.021911621, -0.03817749, 0.0157...

Fig. 1. Example of a multitrack recording from the MUSDB18 dataset

Due to the presence of professionally mixed multitrack recordings, MUSDB18 provides realistic polyphonic audio data. The complexity of overlapping frequency components between different instruments makes this dataset suitable for evaluating separation methods under practical conditions.

The availability of isolated stems for each track enables formulation of the separation task as supervised mask estimation in the time-frequency domain.

Time-Frequency Representation

To perform source separation in the spectral domain, the time-domain audio signals from the MUSDB18 dataset are transformed using the Short-Time Fourier Transform (STFT). Musical signals are non-stationary and contain overlapping spectral components; therefore, localized frequency analysis is required to capture their temporal evolution.

The STFT is computed by applying a window function to short overlapping segments of the signal and performing the Fourier transform within each segment. The mathematical formulation of the transform is given below:

$$X(\tau, \omega) = \int_{-\infty}^{+\infty} x(t) w(t - \tau) e^{-i\omega t} dt \quad (1)$$

The STFT produces a time-frequency representation in which one axis corresponds to time, and the other corresponds to frequency. The magnitude of the complex STFT coefficients forms a spectrogram representation of the signal.

As illustrated in Fig. 2, different musical components manifest distinct structures in the time-frequency domain. Harmonic sources such as vocals or sustained instruments form horizontal patterns corresponding to stable frequency components, while transient components appear as short vertical structures with broad spectral content. This structural diversity enables the formulation of the separation task as a selective manipulation of time-frequency bins.

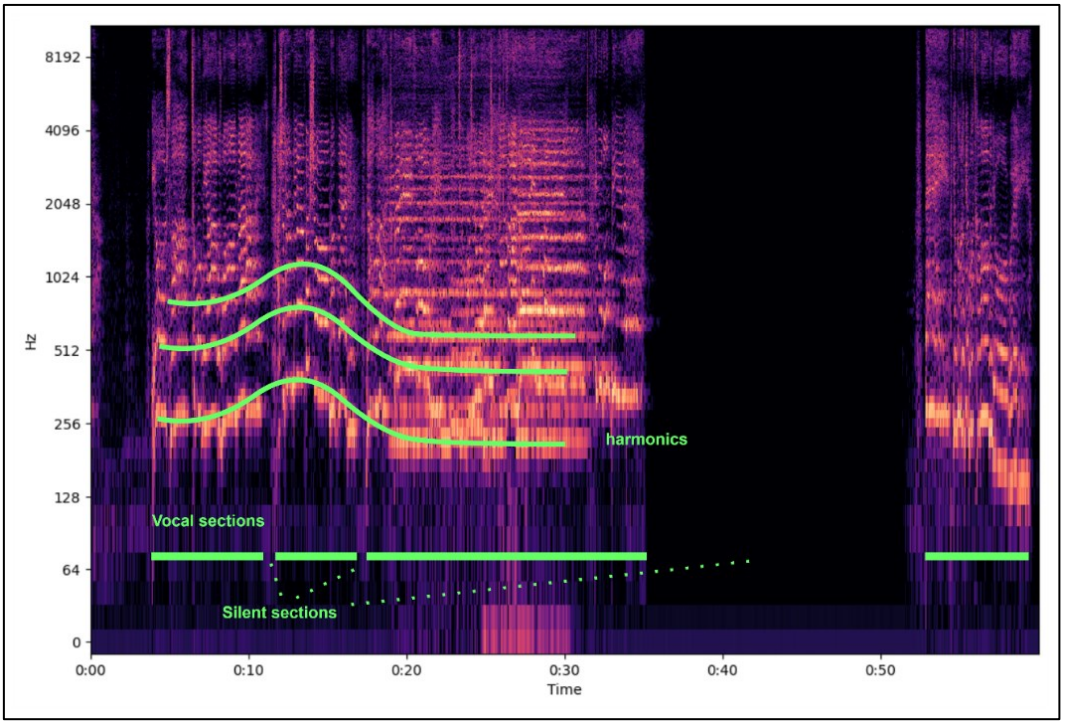

Fig. 2. Example of a spectrogram representation of a mixed musical signal

Prior to transformation, the audio signals are resampled to 22050 Hz. The original recordings sampled at 44100 Hz are downsampled to reduce the dimensionality of the time-frequency representation and improve computational efficiency while preserving the spectral information required for separation. Although this resampling limits the representable frequency range to approximately 11 kHz, the main structural components of musical signals remain preserved. For mask-based separation, the model primarily relies on the overall spectrogram structure rather than on the highest frequency components. As a result, the reduction of the upper frequency range has limited impact on the ability of the model to identify and separate the target sources, although some very high-frequency details, such as cymbal overtones or subtle vocal harmonics, may not be preserved after downsampling.

The STFT is computed using a window size of 1024 samples and a hop size of 256 samples. With this configuration, consecutive windows overlap significantly, providing smooth temporal transitions in the spectral representation. The hop size of 256 samples at a sampling rate of 22050 Hz corresponds to a temporal resolution of approximately 11.6 milliseconds per frame (256 / 22050).

Since the input signals are real-valued, only the non-redundant positive frequency components are retained. For a window size of 1024 samples, this results in 513 frequency bins per time frame.

The STFT produces a complex-valued matrix $S \in \mathbb{C}^{(513 \times T)}$, where 513 corresponds to frequency bins, and T depends on signal duration and hop size. For modeling purposes, only the magnitude of the STFT coefficients is used to form the spectrogram representation, while phase information is preserved for subsequent signal reconstruction.

To incorporate temporal context into the separation process, 25 consecutive STFT frames are grouped together, corresponding to approximately 300 milliseconds of audio. This context window provides sufficient information to capture short musical events while maintaining the temporal localization domain. A context window of 25 frames was used to provide sufficient temporal context for reliable mask estimation while keeping the input size compact and avoiding unnecessary model complexity.

Binary Masking Strategy

After obtaining the time-frequency representation of the mixture signal, the separation task is formulated as the estimation of a binary mask in the spectral domain. Instead of directly predicting the magnitude spectrum of a target source, the model learns to estimate a mask that selects the time-frequency components of that source.

Let $S_m(f,t)$ denote the magnitude spectrogram of the mixture signal, and $S_t(f,t)$ denote the magnitude spectrogram of the corresponding isolated target source. The binary mask is defined as:

$$B(f,t) = \begin{cases} 1, if\ S_t(f,t) > T * S_m(f,t) \\ 0, otherwise \end{cases} \quad (2)$$

Each element of the mask takes a value of 1 if the target source dominates the corresponding time-frequency bin based on the threshold T, and 0 otherwise. While this method is computationally efficient, it can lead to audible artifacts when sources overlap significantly in the same frequency range.

This formulation converts the separation problem into a classification task at the level of individual time-frequency bins. The model does not reconstruct spectral magnitudes directly but predicts whether a particular bin belongs to the target source.

To isolate the desired source, the estimated binary mask $B(f,t)$ is applied element-wise to the magnitude spectrogram of the mixture signal. The estimated target spectrogram $\hat{S}_t(f,t)$ is calculated as:

$$\hat{S}_t(f,t) = B(f,t) * S_m(f,t) \quad (3)$$

The application of this predicted mask suppresses non-target components while preserving the dominant spectral regions of the desired source. (Fig. 3).

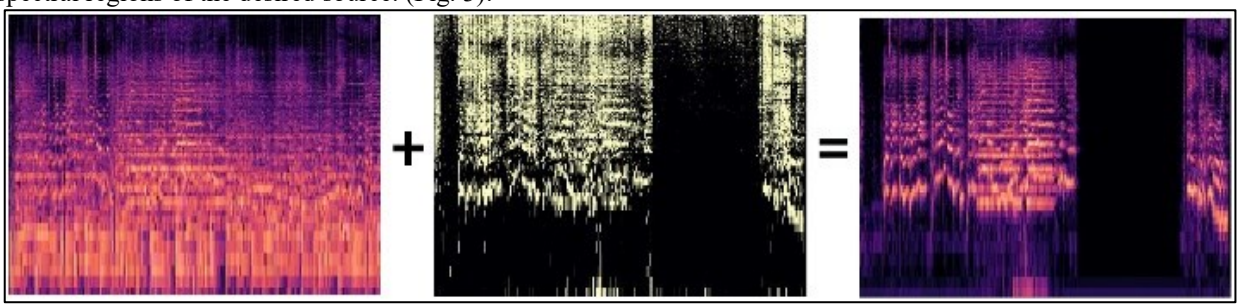

Fig. 3. Example of using a binary mask for vocals separation

Compared to direct spectral regression, binary masking reduces the complexity of the learning task and improves the stability of training, as the network focuses on learning classification boundaries between overlapping sources rather than precise amplitude reconstruction.

Dataset Preparation for Mask Estimation

After defining the binary mask formulation, the dataset is prepared for supervised learning. Each training example consists of a localized time-frequency fragment extracted from the mixture spectrogram and the corresponding fragment of the binary mask computed from the isolated target source.

Let $S_m \in \mathbb{R}^{(513 \times T)}$ denote the magnitude spectrogram of the mixture signal and $B \in \{0,1\}^{(513 \times T)}$ denote the corresponding binary mask. Instead of using the entire spectrogram as a single input, a sliding window approach is applied in the time dimension.

For each time index t, a context window of 25 consecutive frames is extracted:

$$X_t \in \mathbb{R}^{(513 \times 25)}$$

where the central frame corresponds to time index t. The associated target is the mask vector of the central frame:

$$y_t \in \{0,1\}^{(513)}$$

This formulation ensures that the model predicts the mask for a single time frame while taking into account the temporal context obtained from neighboring frames.

The sliding window moves with a step of one frame. As a result, nearly every frame in the spectrogram becomes the center of a training sample. At the beginning and end of the signal, missing frames are handled by replicating the nearest available frame to preserve consistent input dimensions.

For each composition in the training subset, a 60-second segment extracted from the middle of the track is used. This selection avoids fade-in and fade-out regions typically present at the beginning and end of musical recordings, which may introduce low-energy or structurally unrepresentative fragments. Using the central part of each track ensures that the selected segment contains a stable musical structure and active source components. The same duration is applied uniformly to all compositions in order to keep consistent sample generation across the dataset.

Given approximately 5000 STFT frames per 60-second recording, each composition generates around 5000 training samples. Considering 100 compositions in the training subset, the total number of generated samples is approximately 0.5 million.

This input-target pairing converts the separation problem into supervised binary classification at the level of individual frequency bins, where the model is trained to estimate mask values for each spectral component of the central frame.

	MIXTURE_STFT_SPEC_FRAME	MIXTURE_STFT_SPEC	DRUMS_STFT_SPEC	DRUMS_STFT_MASK	BASS_STFT_SPEC	BASS_STFT_MASK	OTHER_STFT_SPEC
0	[[-0.67308205, -0.67308205, -0.67308205, -0.67...	[-0.67308205, 0.7091919, 3.7395122, -9.180677,...	[0.7138199, -1.3688705, 2.3501148, -1.7486998,...	[True, True, True, False, False, False, False,...	[-1.8983451, 2.739884, 0.39843297, -6.092152, ...	[True, True, False, True, True, False, False, ...	[0.7343097 -0.8131887 1.0691648 -1.349947,..
0	[[-0.67308205, -0.67308205, -0.67308205, -0.67...	[-1.5594838, 0.41020006, 2.3798552, -2.7031727...	[0.23175913, 1.05118, -2.9622972, 2.5973017, -...	[False, True, True, True, True, True, False, T...	[-1.792298, -1.0603696, 5.778682, -4.6869626, ...	[True, True, True, True, True, True, Tru...	[0.1550967 0.2580522 -0.382294 -0.4609594.
0	[[-0.67308205, -0.67308205, -0.67308205, -0.67...	[1.3152795, 0.41414323, -10.354743, 13.436728,...	[1.1689886, -1.1697382, 1.0262778, -0.1459337,...	[True, True, False, False, False, False, False...	[0.22298025, 1.7629712, -11.449536, 13.9335985...	[False, True, True, True, True, False, False, ...	[-0.2143786 0.012953159 -0.10878143 -0.20.
0	[[-0.67308205, -0.67308205, -0.67308205, -0.67...	[1.195784, -2.0199337, 8.000207, -8.674113, 7....	[1.686425, -1.1833972, 1.0128316, -0.7019785, ...	[True, False, False, False, False, False, Fals...	[-0.28906628, -1.0251458, 6.878928, -7.867885,...	[False, False, True, True, True, True, False, ...	[-0.2242667 0.1701297 0.1526709 0.104746.
0	[[-0.67308205, -0.67308205, -0.67308205, -0.67...	[0.516275, 0.009941088, 2.8034518, -4.7309585,...	[1.250725, 0.013183201, -1.5002512, 1.9060094,...	[True, True, False, False, True, False, False,...	[-0.28617603, -0.2626148, 4.306538, -6.445273,...	[False, True, True, True, True, True, Tr...	[-0.3592853 0.1939049 -0.0230061 -0.088.
...	...	...	...	...	...	...	.
49	[[2.883418, -0.09167261, 3.1611714, -1.7571845...	[1.0679209, -1.948537, 2.5397038, 0.7162589, -...	[0.50335455, -1.3976977, 0.59142345, 4.1875033...	[False, True, False, True, True, False, False,...	[0.003926173, -0.020545062, 1.579238, -3.24491...	[False, False, True, True, True, False,...	[-0.0683219 -0.0281803 0.092211984, -0.05.
49	[[-0.09167261, 3.1611714, -1.7571845, -2.03464...	[0.912037, 1.0700151, -0.73040056, -2.9615598,...	[0.47293648, 1.3671837, -0.26324585, -4.040459...	[False, True, False, True, True, False, False,...	[0.15331633, -0.18774748, -0.3747006, 0.965761...	[False, False, False, False, False, False, Fal...	[0.026956 0.000580 -0.0107219815, 0.0.
49	[[3.1611714, -1.7571845, -2.0346498, 0.8381055...	[1.0327331, -1.7265631, 3.3210046, -4.4533873,...	[1.3227125, -1.9820887, 5.0938287, -7.856887, ...	[True, True, True, True, True, False, False, F...	[-0.009553965, 0.06717502, -1.7113749, 3.39441...	[False, False, False, True, True, False, False...	[0.014159967 8.710668e-0 -0.0060033, -0.085.
49	[[-1.7571845, -2.0346498, 0.83810556, 0.296058...	[-0.7497989, 1.0227028, -1.3939776, 1.3733646,...	[-0.2605092, 0.83635867, -1.7464806, 1.6133298...	[False, True, True, True, True, False, False, ...	[0.11254756, -0.13039157, 0.22599964, -0.20880...	[False, False, False, False, False, False, Fal...	[-0.041799 -0.024618 0.16930972, -0.05.
49	[[-2.0346498, 0.83810556, 0.29605883, 2.441881...	[-1.4421206, 1.1397387, -1.8002272, 4.4827785,...	[-0.17732884, 1.0233698, -3.8529348, 6.303258,...	[False, True, True, True, True, False, False, ...	[-0.70629275, -0.1407858, 2.3236382, -2.853734...	[False, False, True, True, False, False, False...	[-0.375910 0.013332 0.054240 0.93358.

Fig. 4. Structure of the processed dataset with frame-based STFT features of the mixture and sources and their corresponding binary masks

No explicit data augmentation techniques such as tempo modification, pitch shifting, or artificial noise addition were applied during training. The MUSDB18 dataset already contains a wide variety of musical styles and instrument combinations, which provides sufficient variability for training and reduces the risk of severe overfitting. In this study, the experimental setup focuses on evaluating the separation capability of the proposed architecture under controlled conditions rather than introducing additional variability through data augmentation. Additionally, regularization techniques such as dropout were used within the network architecture to further improve generalization.

Network Architecture and Signal Representation

To estimate the binary mask from localized spectrogram fragments, a convolutional neural network architecture is employed. The model is designed to process two-dimensional time-frequency representations of size 513x25, where 513 corresponds to frequency bins, and 25 corresponds to consecutive time frames.

Table 1.

Detailed layer specifications of the CNN architecture for sound source separation

Layer Type	Kernel	Stride	Padding	Output Shape	Number of Parameters
Input spectrogram segment	--	--	--	1x513x25	--
Conv2D (32 filters) + LeakyReLU	3x3	1x1	same	32x513x25	320
Conv2D (16 filters) + LeakyReLU	3x3	1x1	same	16x513x25	4,624
MaxPooling2D + Dropout (0.1)	3x3	3x3	0	16x171x8	--
Conv2D (64 filters) + LeakyReLU	3x3	1x1	same	64x171x8	9,280
Conv2D (16 filters) + LeakyReLU	3x3	1x1	same	16x171x8	9,232
MaxPooling2D + Dropout (0.1)	3x3	3x3	0	16x57x2	--
Flatten	--	--	--	1824	--
Fully Connected (1824 > 128) + LeakyReLU + Dropout (0.2)	--	--	--	128	233,600
Fully Connected (128 > 513) + Sigmoid	--	--	--	513	66,177
Total					323,233

The architecture consists of multiple convolutional layers followed by nonlinear activation functions and fully connected layers that produce the final mask prediction for the central frame.

The convolutional layers extract localized spectral-temporal features by applying learnable filters across both frequency and time dimensions. This allows the network to capture structured patterns such as harmonic stacks, transient bursts, and broadband noise components that characterize different musical sources.

Table 1 illustrates the layer specifications of the convolutional neural network used for mask estimation. The intermediate feature maps generated by the convolutional layers represent progressively abstract representations of the input spectrogram fragment. Lower layers focus on local spectral structures, while deeper layers capture higher-level combinations of time-frequency patterns.

Fig. 5. presents the architecture of a Convolutional Neural Network for sound source separation. The diagram visualizes the transformation of intermediate feature maps during forward propagation, showing the progression from the input spectrogram through convolutional layers to the final fully connected layers (dimensions 1024, 128, and 513).

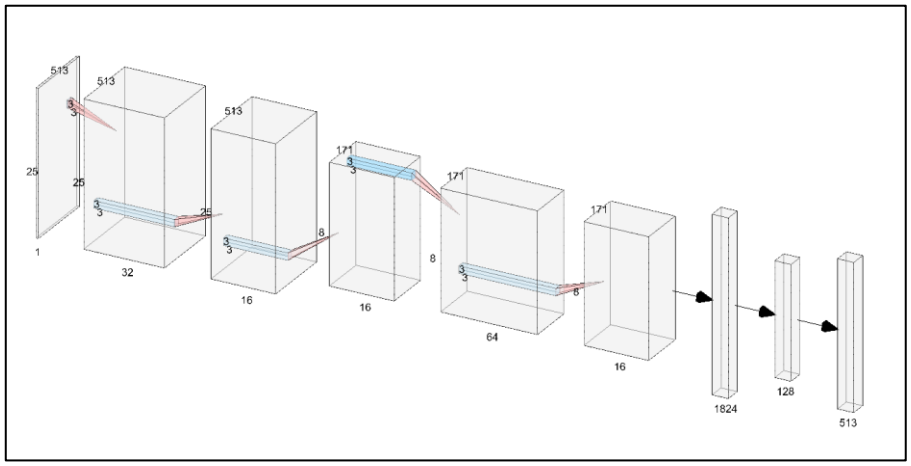

Fig. 5. General architecture of a CNN for sound source separation

The hyperparameters of the network were selected empirically to balance model performance and architectural simplicity. In particular, parameters such as the number of convolutional filters, kernel sizes, and the size of the fully connected layer affect the amount of information the model can capture and the required computational resources. Increasing these parameters may improve separation accuracy but also increases computational cost and model size.

After convolutional processing, the feature maps are flattened and passed to fully connected layers, which aggregate extracted features and produce an output vector of dimension 513. Each output element corresponds to a predicted mask value for a specific frequency bin of the central frame.

A sigmoid activation function is applied at the output layer to constrain predictions to the range [0,1], consistent with binary mask formulation. Fig. 6 demonstrates the predicted mask output produced by the trained CNN.

An implementation of the proposed model and the experimental pipeline used in this study is available in an open repository [5].

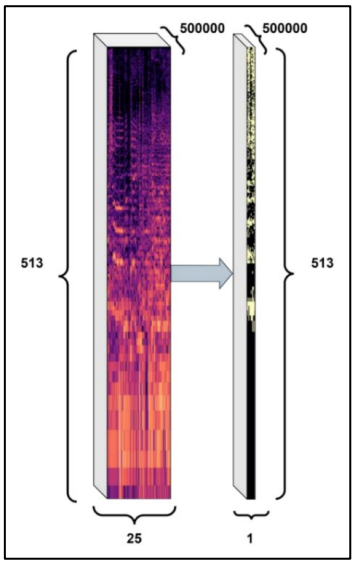

Fig. 6. Binary mask prediction scheme for a specific sound source

Research results

The proposed convolutional neural network model for binary mask estimation was trained on 100 musical compositions from the MUSDB18 dataset. For each composition, a 60-second segment extracted from the central part of the track was used to avoid fade-in and fade-out regions. The remaining 50 compositions were reserved exclusively for evaluation.

Training was performed for 50 epochs using Stochastic Gradient Descent (SGD) as the optimization algorithm. SGD was selected due to its lower memory overhead compared to adaptive optimizers and its suitability for training on large datasets containing heterogeneous audio sources. In the proposed study, the same neural network architecture is trained for multiple source categories, including vocals, drums, bass, and other instruments. This requires the model to generalize across different spectral structures while maintaining stable optimization during training.

The network was trained using the Mean Squared Error (MSE) loss between the predicted mask and the ground-truth mask. Although the final mask is interpreted as a binary classification of time-frequency bins, the learning problem is formulated as a regression task. In this formulation, the objective of the model is to approximate the ideal binary mask rather than to directly reconstruct the time-domain signal. Consequently, the training and evaluation primarily focus on mask estimation accuracy rather than on commonly used waveform reconstruction metrics. For binary mask estimation tasks, metrics such as MSE and Dice coefficient are commonly used to quantify the similarity between the predicted mask and the ideal binary mask. These measures directly reflect the accuracy of time-frequency bin assignment and therefore provide an informative evaluation of mask reconstruction quality.

To improve convergence stability and avoid premature stagnation, a cyclic learning rate scheduling strategy (CyclicLR) was applied during training under a triangular policy. During reconstruction, the predicted data was converted into a binary mask using a fixed threshold $T = 0.6$. This value was chosen empirically as a sufficiently low yet robust threshold that captures most dominant components of the target source while minimizing interference from other sources.

The performance of the proposed convolutional neural network model was evaluated separately for each target sound source. Training was conducted for 50 epochs using the prepared dataset derived from MUSDB18. The quantitative results are summarized in Table 2.

Table 2.

Accuracy and loss results for each source

Source type	Dataset type	Accuracy	Loss value (MSE)
Vocal isolation	Training	0.772	0.153
	Validation	0.743	0.174
Drum isolation	Training	0.766	0.159
	Validation	0.685	0.203
Bass isolation	Training	0.944	0.044
	Validation	0.939	0.051
Other source isolation	Training	0.764	0.158
	Validation	0.732	0.1791

For vocal isolation, the model achieved a training accuracy of 0.772 and a validation accuracy of 0.743, with corresponding loss values of 0.153 and 0.174. The training and validation accuracy and loss curves for the vocal binary mask prediction network are presented in Fig. 7.

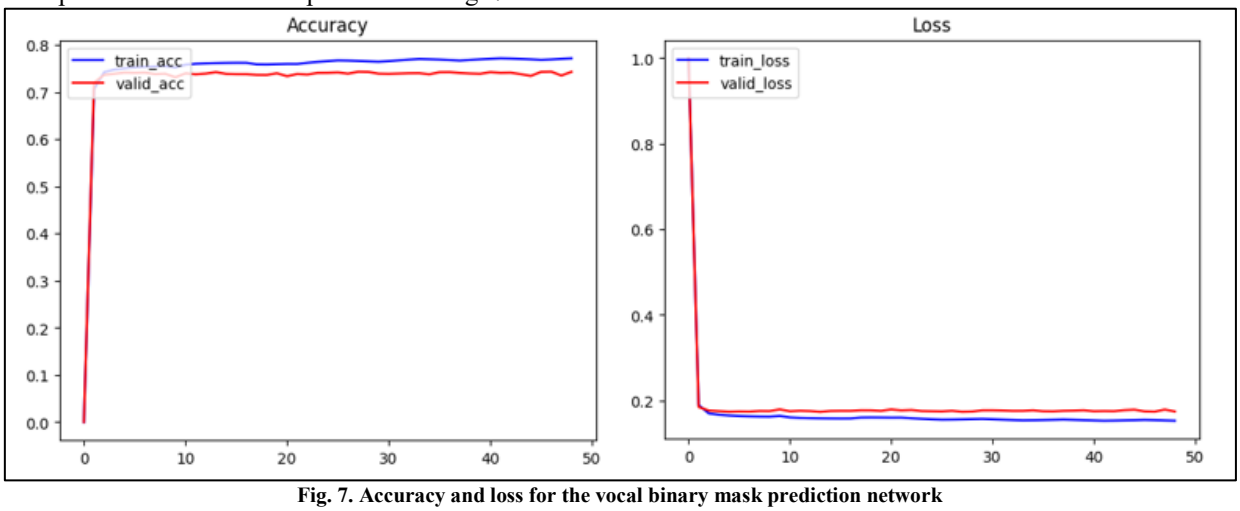

Fig. 7. Accuracy and loss for the vocal binary mask prediction network

As shown in Fig. 7, both accuracy curves increase during training and stabilize toward the final epochs. The loss function decreases consistently for both datasets. A comparison between the predicted vocal mask and the ground

truth mask is shown in Fig. 8. The predicted mask reproduces the main time-frequency regions corresponding to vocal components.

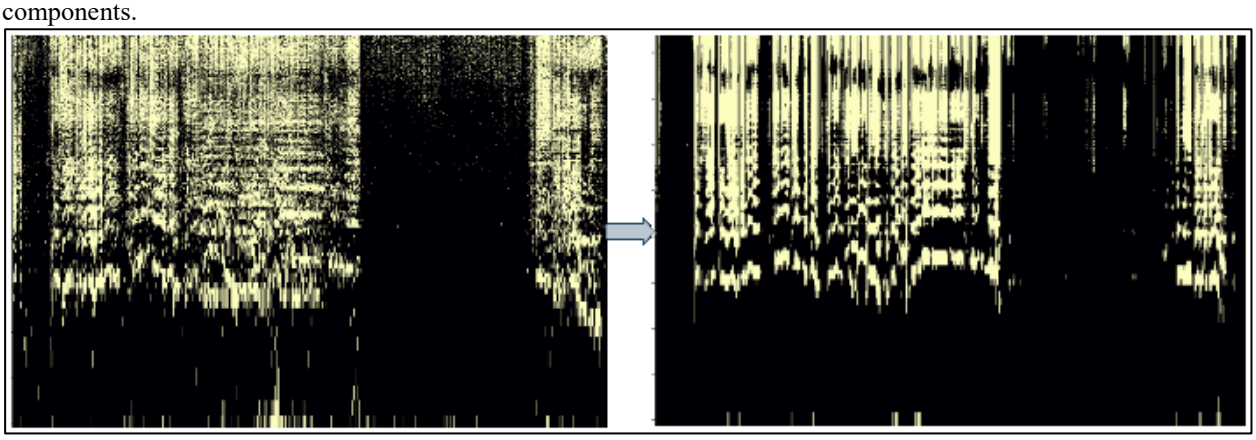

Fig. 8. Comparison of the resulting vocal binary mask with the ground truth

For drums isolation, the model achieved a training accuracy of 0.766 and a validation accuracy of 0.685. The corresponding loss values were 0.159 and 0.203.

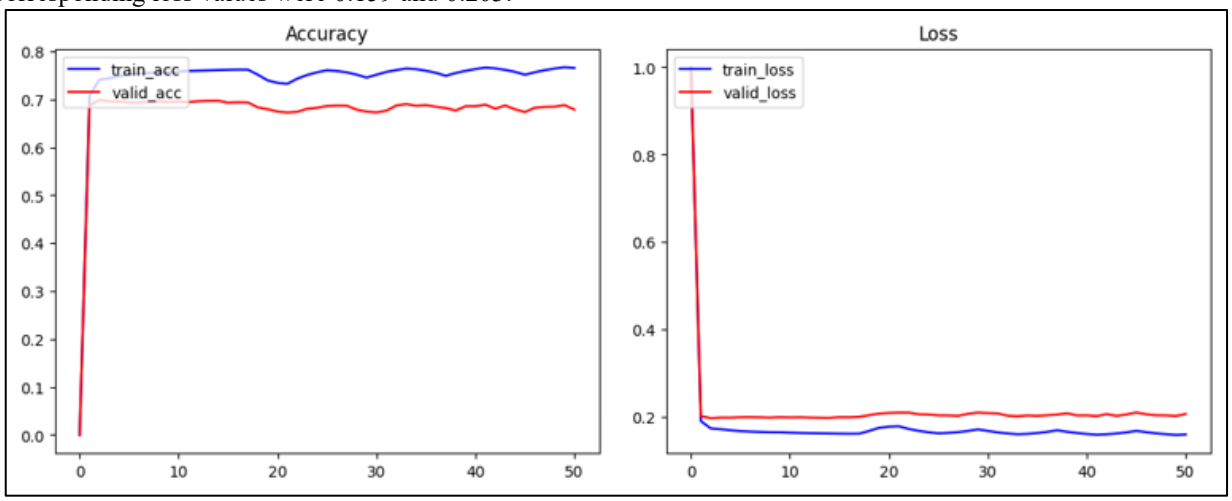

Fig. 9. Accuracy and loss for the drums binary mask prediction network

The training process for drum mask prediction is illustrated in Fig. 9. The curves demonstrate improvement in accuracy and reduction in loss over the course of training.

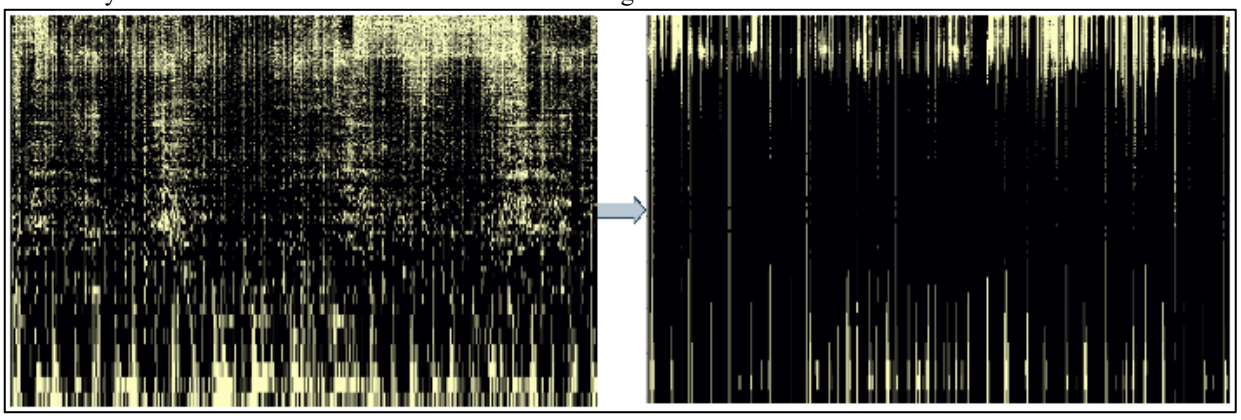

Fig. 10. Comparison of the resulting drums binary mask with the ground truth

An example of the predicted drum mask compared to the ground truth is presented in Fig. 10. The predicted mask captures the principal transient regions associated with drum components.

Bass isolation demonstrated the highest performance, with a training accuracy of 0.944 and a validation accuracy of 0.939. The loss values were 0.044 and 0.051 respectively.

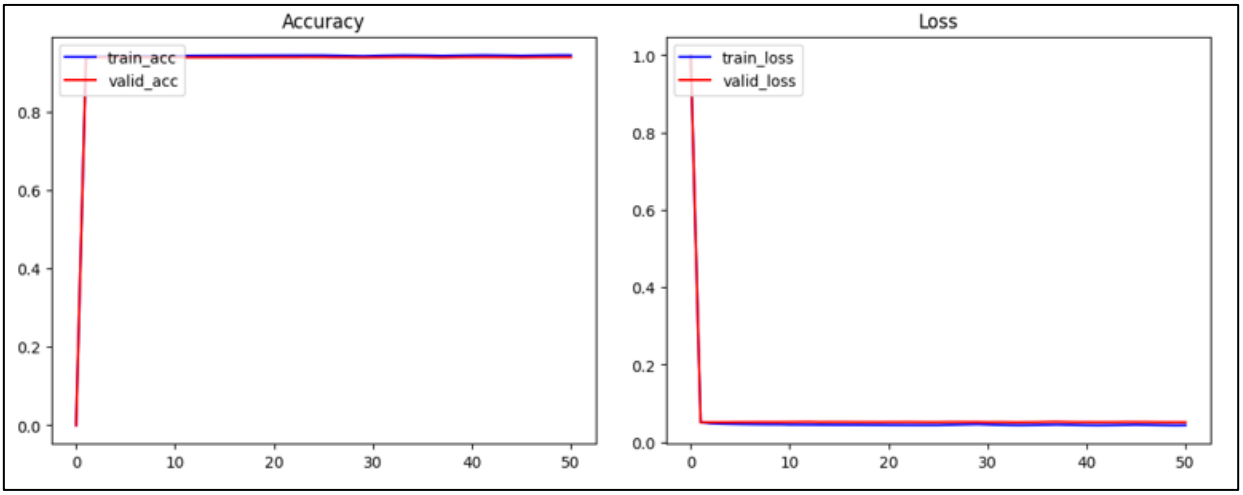

Fig. 11. Accuracy and loss for the bass binary mask prediction network

The corresponding training and validation curves are shown in Fig. 11. The accuracy rapidly increases in early epochs and remains stable during later stages of training.

Fig. 12 presents a comparison between predicted and ground truth bass masks. The predicted mask accurately reflects the dominant low-frequency regions of the target source.

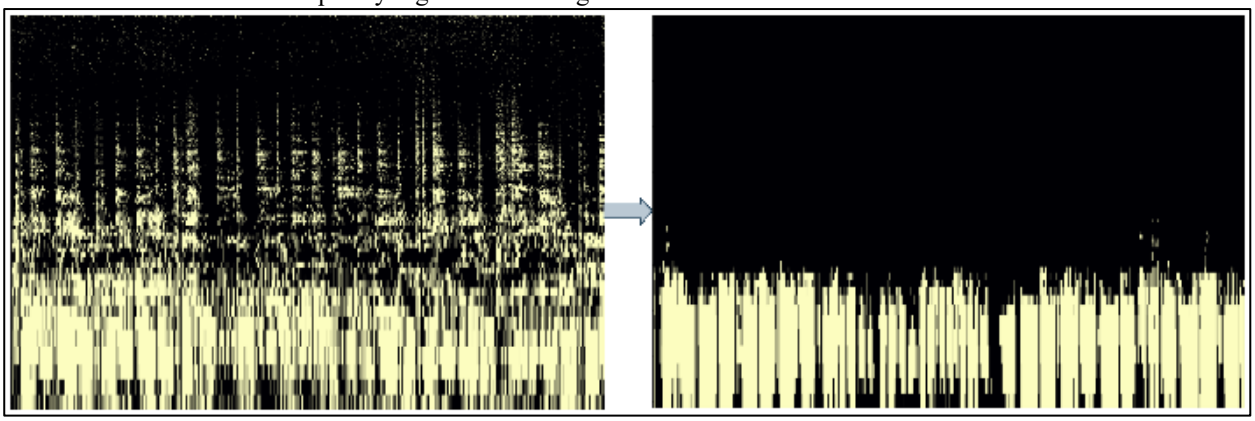

Fig. 12. Comparison of the resulting bass binary mask with the ground truth

For the isolation of the ‘other’ category, the model achieved a training accuracy of 0.764 and a validation accuracy of 0.732. The loss values were 0.158 and 0.1791.

The training dynamics for this source category are shown in Fig. 13. The curves demonstrate stable convergence behavior.

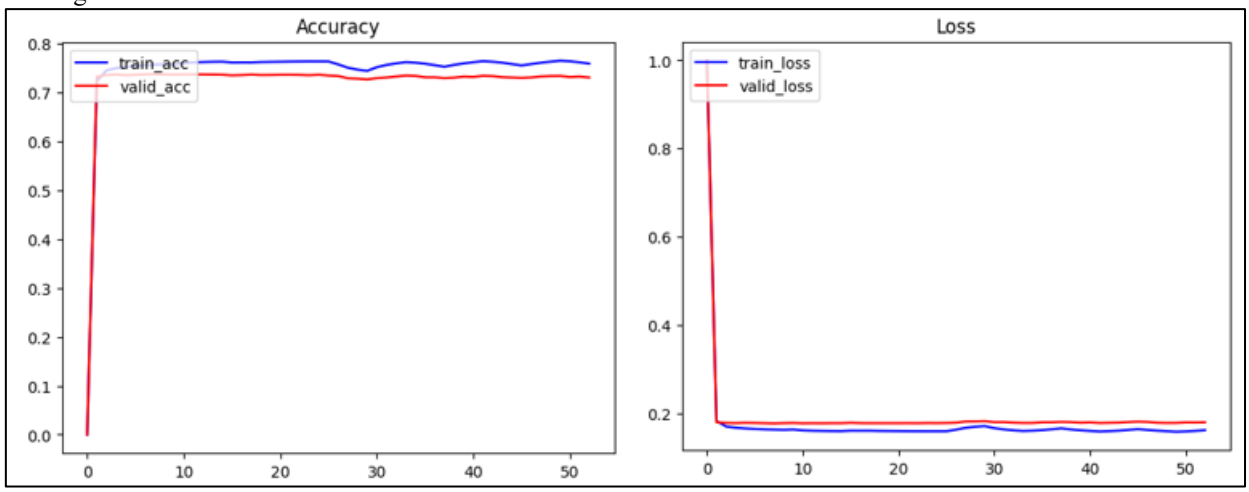

Fig. 13. Accuracy and loss for the ‘other’ audio binary mask prediction network

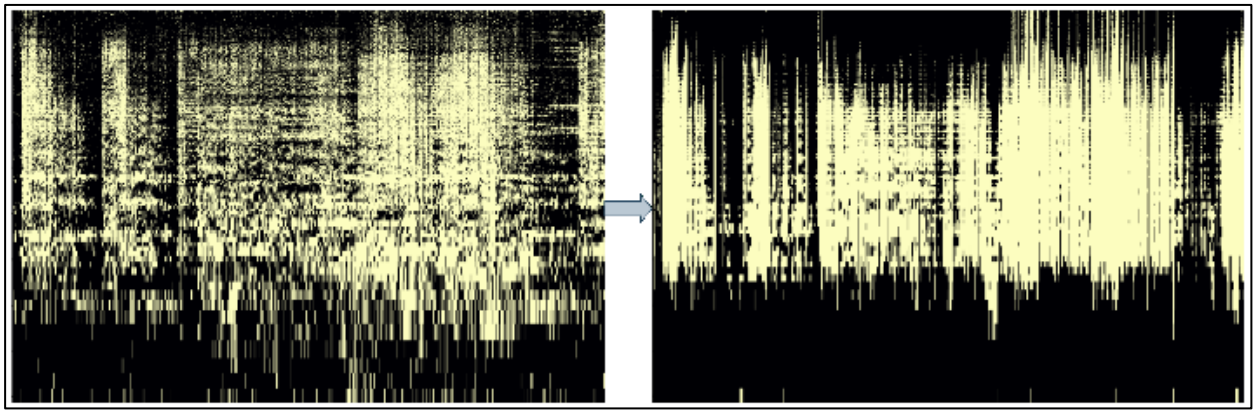

Fig. 14. Comparison of the resulting 'other' audio binary mask with the ground truth

Fig. 14 illustrates the comparison between the predicted and ground truth masks for 'other' sources. The predicted mask identifies the main spectral regions associated with instrumental accompaniment.

To further quantify the similarity between the predicted and ideal binary masks, the Dice coefficient was additionally computed over all processed samples. The obtained values are summarized in Table 3.

Table 3.

Dice coefficient for predicted binary masks

Source type	Vocal	Drum	Bass	Other source
Mean Dice coefficient	0.697659	0.734685	0.599432	0.679861

The Dice coefficient provides a measure of the overlap between binary mask regions in the time-frequency representation. In contrast to point-wise error metrics such as MSE, it evaluates how consistently the model assigns time-frequency bins to the corresponding sources. Therefore, it offers a complementary perspective on the quality of mask estimation.

Along with mask-based metrics, several signal-level indicators were evaluated for the reconstructed from predicted masks audio signals, including SI-SDR, SDR, and SNR. These measures characterize the amount of distortion and residual interference present in the separated signals after reconstruction. The obtained values demonstrated in Table 4 indicate that the proposed approach produces perceptually meaningful separation across different source types, while maintaining computational efficiency of the model.

Table 4

Signal-level evaluation metrics for reconstructed sources

Source type	**SI-SDR (dB)**	**SDR (dB)**	**SNR (dB)**
Vocal	3.222489	4.532568	5.094826
Drums	-0.960971	1.016916	3.307660
Bass	4.101974	4.592130	5.012954
Other	-1.237303	-0.257126	1.159419

In addition to the reported separation results, the proposed architecture demonstrates low computational requirements. Under the experimental conditions, processing a one-minute audio segment requires approximately 3 seconds, corresponding to a processing rate of about twenty seconds of audio per second of computation.

Overall, the obtained results demonstrate consistent separation performance across different sound sources. The reported accuracy and Dice coefficient values show stable binary mask estimation and satisfactory reconstruction quality for the evaluated source categories.

Discussion of results

To provide an overall assessment of the proposed model performance, the averaged quantitative results across all isolated sources are summarized in Table 5. The mean training accuracy reached 0.811, while the corresponding validation accuracy was 0.775. The difference of approximately 3.6 percent indicates controlled generalization behavior without significant overfitting.

Table 5.

Averaged accuracy and loss results

Dataset type	**Accuracy**	**Loss value (MSE)**
Training	0.811	0.129
Validation	0.775	0.152

The highest separation performance was obtained for bass isolation, where accuracy approached 0.94. This exceeds the averaged validation performance and confirms that low-frequency harmonic components are effectively captured by the convolutional architecture. The higher performance observed for bass separation can be explained by its relatively narrow spectral range and more stable harmonic structure compared to other sources. Bass components are typically concentrated in the low-frequency region, where spectral overlap with other instruments is reduced, making them easier to isolate using spectrogram-based methods.

Vocal isolation achieved a validation accuracy of 0.743, which is approximately 19 percent lower than bass performance. This difference reflects the increased spectral overlap of vocal components with mid-frequency instrumental content.

The performance obtained for the ‘other’ category reaches a validation accuracy of 0.732, which is comparable to the vocal separation result of 0.743 and represents an intermediate level among the evaluated sources. This result can be explained by the heterogeneous composition of this category. Unlike vocals or bass, the ‘other’ class includes a wide range of instruments with different spectral and temporal characteristics.

Drum isolation demonstrated the lowest validation accuracy of 0.685. Compared to bass isolation, the performance difference exceeds 25 percent. The transient and broadband nature of percussive signals significantly increases ambiguity in binary mask assignment, especially in dense mixtures. Although, unlike the heterogeneous ‘other’ class, drum signals correspond to a single source category, they often contain a combination of strong low-frequency components from kick and tom drums together with high-frequency transients from cymbals, producing sharp broadband spikes in the spectrogram that are more difficult for the model to represent consistently.

The gap between training and validation accuracy remains moderate for all sources. For vocals, the difference between training (0.772) and validation (0.743) accuracy is approximately 2.9 percent. For drums, the difference reaches approximately 8 percent, indicating slightly higher sensitivity to unseen data in percussive modeling.

As shown in Table 3, the Dice coefficient values provide additional insight into the spatial consistency of the predicted masks. The highest Dice score was obtained for drums separation (0.735), indicating that the predicted masks capture the main transient regions associated with percussive components. Vocal and ‘other’ sources demonstrate comparable Dice values of 0.698 and 0.68 respectively, which may be related to the fact that both groups often occupy similar mid-frequency regions in the time-frequency representation. Bass separation achieved a Dice coefficient of 0.599, which is lower than for the previously discussed evaluated sources.

This result differs from the accuracy-based evaluation reported earlier, where bass achieved the highest accuracy. The difference can be explained by the nature of the evaluated metrics. Accuracy evaluates individual time-frequency bins independently and is therefore strongly influenced by the large number of background bins that are correctly classified. In contrast, the Dice coefficient evaluates the spatial overlap between predicted and reference mask regions. As illustrated in Fig. 12, the predicted bass mask captures the dominant low-frequency region but appears more compact than the corresponding ground-truth mask. This suggests that the model tends to suppress weaker time-frequency components and background regions, which reduces the amount of residual noise in the predicted mask. However, because the Dice metric requires precise overlap between mask regions, excluding such peripheral bins can reduce the Dice score even when the main spectral region of the source is correctly detected.

Signal-level evaluation metrics were computed to assess the quality of the reconstructed audio signals. The obtained values for SI-SDR, SDR, and SNR are summarized in Table 4. These indicators characterize different aspects of reconstruction quality: SI-SDR reflects distortion of the reconstructed signal after scale normalization, SDR describes the overall deviation from the reference signal, and SNR represents the ratio between the target signal energy and residual noise components.

Bass separation demonstrates the strongest reconstruction quality. The SI-SDR value reaches 4.10 dB, indicating that the dominant structure of the bass signal is preserved with relatively low distortion. The corresponding SDR value of 4.59 dB shows that the reconstructed signal remains close to the reference signal, while the SNR value of 5.01 dB indicates that the main low-frequency energy is recovered with limited background interference.

Vocal audio demonstrates comparable behavior. The SI-SDR value of 3.22 dB indicates relatively low reconstruction distortion, while the SDR value of 4.53 dB confirms that the overall signal structure is preserved effectively. The SNR value reaches 5.09 dB, suggesting that the dominant vocal components remain clearly distinguishable from residual noise.

For drums audio, the obtained metrics are noticeably lower. The SI-SDR value of -0.96 dB reflects increased distortion of the reconstructed signal, while the SDR value of 1.02 dB indicates a lower similarity to the reference signal compared to bass and vocals. The SNR value of 3.31 dB further suggests that residual noise components are more pronounced in the reconstructed signal.

The ‘other’ source category demonstrates the lowest reconstruction metrics. The SI-SDR value reaches -1.24 dB, while the SDR value is close to zero at -0.26 dB, indicating lower degree of signal correspondence between the reconstructed and reference signals. The corresponding SNR value of 1.16 dB suggests a higher level of residual interference, which can be explained by the heterogeneous composition of this source category that includes multiple instruments with diverse spectral characteristics.

The binary mask formulation contributes to stable convergence, as reflected in consistent loss reduction during training. However, strict binary decisions may introduce reconstruction artifacts in frequency regions where multiple sources contribute comparable energy.

Additionally, since phase information is not explicitly estimated and the mixture phase is reused during reconstruction, residual distortions may remain even when magnitude prediction is accurate. Estimating phase directly would require neural architectures capable of operating on the complex-valued representation of the STFT, which substantially increases model complexity and introduces additional challenges during training. Therefore, the proposed approach focuses on magnitude estimation while reusing the mixture phase during reconstruction, since the magnitude component typically contains the dominant structural information of the signal in the time-frequency domain. This simplification reduces model complexity while still allowing reliable source separation, although it may introduce minor reconstruction artifacts in the resulting signal.

Overall, the numerical results confirm that the proposed model achieves consistent performance across source types, with accuracy ranging from 0.685 to approximately 0.94 depending on spectral characteristics.

Conclusions

In this work, the task was to isolate different audio signal sources, in particular, vocals, drums, bass, and ‘other’ components. The hypothesis was that the use of convolutional neural networks would allow for the effective determination of the presence or absence of each sound source in the input spectrogram and separate them.

To solve this problem, a Short-Time Fourier transform was used, which allowed us to represent audio signals in the time-frequency domain. Then, convolutional neural networks were used to process the resulting spectrograms. The use of binary masks instead of direct STFT made it possible to simplify the problem and change the approach to a hybrid one, combining elements of regression and classification. The use of CNN allowed the model to detect acoustic patterns and properties on the input spectrograms corresponding to different sound sources.

The results showed high accuracy rates for isolating different audio sources. In particular, the accuracy for isolating vocals was about 77.2%, for bass – about 94.4%, for drums – about 76.6%, and for ‘other’ sources – about 76.4%. In addition, the mean square error was also quite low on the training and validation datasets, averaging about 0.13 for the training and 0.15 for the validation sets.

Evaluation of mask overlap using the Dice coefficient confirms the stability of the predicted time-frequency regions. The highest Dice score was observed for drum signals (0.735), indicating that the model accurately captures the main spectral patterns of this source type. Vocal and ‘other’ sources produced comparable overlap values of 0.698 and 0.68, while bass masks demonstrated a lower Dice value of 0.599.

Signal-level evaluation metrics further confirm the effectiveness of the proposed approach. For bass signals, the model achieved an SI-SDR of 4.10 dB, an SDR of 4.59 dB, and an SNR of 5.01 dB, indicating stable reconstruction of the dominant low-frequency components. Vocal separation produced comparable results, with an SI-SDR of 3.22 dB, an SDR of 4.53 dB, and an SNR of 5.09 dB, suggesting reliable preservation of the main spectral structure of the signal. Lower reconstruction quality was observed for drum signals, where SI-SDR reached -0.96 dB, SDR 1.02 dB, and SNR 3.31 dB, reflecting the increased difficulty of reconstructing transient percussion components. The ‘other’ source category demonstrated the lowest results, with SI-SDR of -1.24 dB, SDR of -0.26 dB, and SNR of 1.16 dB, which can be attributed to the heterogeneous composition of this group of instruments.

In summary, the obtained results indicate that the proposed approach provides a viable solution for separating multiple sound sources in complex audio mixtures. It demonstrates that binary mask prediction based on spectrogram representations can serve as a structurally simple yet effective alternative to more complex end-to-end architectures. At the same time, the observed performance differences between different sound sources highlight remaining limitations related to spectral overlap and mask ambiguity in dense mixtures. The results confirm the viability of the proposed framework and provide a basis for further improvement and extension.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

The authors' contributions are as follows: O. Tkachuk proposed the methodology, designed the algorithms, and carried out the software implementation; O. Tomashevskyy supervised the study and determined the results validation strategy.

DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

Authors acknowledge the use of artificial intelligence tools for assistance in English language translation and formatting, and Grammarly to check the grammar and spelling. The scientific concept, architectural design, implementation, and final text verification remain the sole responsibility of the authors.

REFERENCES

1. Acoustica | Digital Audio Editor. URL: https://acondigital.com/products/acoustica (дата звернення: 11.11.2023).

2. Bhattarai B., Pandeya Y. R., Jie Y. та ін. High-Resolution Representation Learning and Recurrent Neural Network for Singing Voice Separation. *Circuits, Systems, and Signal Processing*. Вип. 42, № 2. С. 1083—1104. DOI:10.1007/s00034-022-02166-5.

3. Cho J., Yoo C. D. Underdetermined Convolutive BSS: Bayes Risk Minimization Based on a Mi xture of Super-Gaussian Posterior Approximation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Вип. 23, № 5. С. 828—839. DOI:10.1109/TASLP.2015.2409778.

4. Du J., Tu Y., Dai L.-R. та ін. A Regression Approach to Single-Channel Speech Separation Via High-Res olution Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Вип. 24, № 8. С. 1424—1437. DOI:10.1109/TASLP.2016.2558822.

5. Edwin1349/masters [Електронний ресурс] – Режим доступу https://github.com/Edwin1349/masters. *GitHub*. URL: https://github.com/Edwin1349/masters (дата звернення: 12.03.2026).

6. León J. P., Beltrán J. R. Blind separation of overlapping partials in harmonic musical notes usi ng amplitude and phase reconstruction. *EURASIP Journal on Advances in Signal Processing*. Вип. 2012, № 1. С. 223. DOI:10.1186/1687-6180-2012-223.

7. Mezza A. I., Giampiccolo R., Bernardini A. та ін. Toward deep drum source separation. *Pattern Recognition Letters*. Вип. 183, 07.2024. С. 86—91. DOI:10.1016/j.patrec.2024.04.026.

8. Mika D., Budzik G., Józwik J. Single Channel Source Separation with ICA-Based Time-Frequency Decompo sition. *Sensors*. Вип. 20, № 7. С. 2019. DOI:10.3390/s20072019.

9. SpectraLayers: Advanced Spectral Audio Editor [Електронний ресурс] – Р ежим доступу https://new.steinberg.net/spectralayers/. URL: https://www.steinberg.net/spectralayers/ (дата звернення: 11.11.2023).

10. Virtanen T., Gemmeke J. F., Raj B. Active-Set Newton Algorithm for Overcomplete Non-Negative Representati ons of Audio. *IEEE Transactions on Audio, Speech, and Language Processing*. Вип. 21, № 11. С. 2277—2289. DOI:10.1109/TASL.2013.2263144.

11. VirtualDJ - Real-Time Stems Separation [Електронний ресурс] – Режим до ступу https://www.virtualdj.com/stems/. URL: https://www.virtualdj.com/stems/ (дата звернення: 11.11.2023).

12. Woo W. L., Dlay S. S., Al-Tmeme A. та ін. Reverberant signal separation using optimized complex sparse nonnegative tensor deconvolution on spectral covariance matrix. *Digital Signal Processing*. Вип. 83, 12.2018. С. 9—23. DOI:10.1016/j.dsp.2018.07.018.

13. Ansari S., Alnajjar K. A., Khater T. et al. A Robust Hybrid Neural Network Architecture for Blind Source Separation of Speech Signals Exploiting Deep Learning. *IEEE Access*. Vol. 11, 2023. P. 100414—100437. DOI:10.1109/ACCESS.2023.3313972.

14. DJ Apps | Algoriddim [Електронний ресурс] – Режим доступу https://www. algoriddim.com/apps. URL: https://www.algoriddim.com/apps (accessed 11.11.2023).

15. Feng Y. Single and Multichannel Speech Source Separation using Non- Negative M atrix Factorisation Incorporating Spectral Masks. *University of Wollongong Thesis Collection 2017*. 01.01.2017. URL: https://ro.uow.edu.au/theses1/90

16. Grais E. M., Roma G., Simpson A. J. R. et al. Two-Stage Single-Channel Audio Source Separation Using Deep Neural Net works. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Vol. 25, Issue 9. P. 1773—1783. DOI:10.1109/TASLP.2017.2716443.

17. Hennequin R., Khlif A., Voituret F. et al. Spleeter: a fast and efficient music source separation tool with pre-t rained models. *Journal of Open Source Software*. Vol. 5, Issue 50. P. 2154. DOI:10.21105/joss.02154.

18. Koldovskỳ Z., Málek J., Janskỳ J. Extraction of independent vector component from underdetermined mixtures through block-wise determined modeling. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech an d Signal Processing (ICASSP)*(2019). IEEE, 2019. DOI:10.1109/ICASSP.2019.8683431. P. 7903—7907.

19. Leplat V., Gillis N., Ang A. M. S. Blind Audio Source Separation With Minimum-Volume Beta-Divergence NMF. *IEEE Transactions on Signal Processing*. Vol. 68, 2020. P. 3400—3410. DOI:10.1109/TSP.2020.2991801.

20. Luo Y., Mesgarani N. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Spe ech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Vol. 27, Issue 8. P. 1256—1266. DOI:10.1109/TASLP.2019.2915167.

21. Rafii Z., Liutkus A., Stöter F.-R. et al. MUSDB18 - a corpus for music separation. (17.12.2017). Zenodo, 2017. DOI:10.5281/ZENODO.1117372. 2017.

22. Shop RX 10 Standard | iZotope [Електронний ресурс] – Режим доступу htt ps://www.izotope.com/en/shop/rx-8-standard.html. URL: https://www.izotope.com/en/shop/rx-10-standard/ (accessed 11.11.2023).

23. Zhan G., Huang Z., Ying D. et al. Improvement of mask-based speech source separation

using DNN. *2016 10th International Symposium on Chinese Spoken Language Processin g (ISCSLP)*(2016). IEEE, 2016. P. 1—5.

24. Zhang L., Li C., Deng F. et al. Multi-Task Audio Source Separation. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASR U)*(13.12.2021). Cartagena, Colombia : IEEE, 2021. DOI:10.1109/ASRU51503.2021.9687922. P. 671—678.

Олег ТОМАШЕВСЬКИЙ, Орест ТКАЧУК
Національний університет «Львівська політехніка»

РОЗДІЛЕННЯ ЗВУКОВИХ ДЖЕРЕЛ У ЧАСОВО-ЧАСТОТНІЙ ОБЛАСТІ НА ОСНОВІ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ

У роботі розглядається задача розділення звукових джерел у змішаних аудіосигналах у часово-частотній області. Досліджується застосування згорткових нейронних мереж для ізоляції окремих акустичних компонентів зі складних аудіоміксів, у яких декілька джерел перекриваються як у часі, так і за частотою. Наявність такого перекриття суттєво ускладнює процес розділення та підвищує вимоги до стабільності й структурної узгодженості застосованих моделей. Запропонований підхід базується на перетворенні аудіосигналів за допомогою віконного перетворення Фур'є та поданні аудіоміксів у вигляді спектрограм, що зберігають як часові, так і спектральні характеристики звукових компонентів. До отриманих представлень застосовується стратегія бінарного маскування з метою структурного спрощення задачі розділення. Згорткова нейронна мережа використовується для прогнозування масок, що відповідають окремим звуковим джерелам, таким як вокал, бас, барабани та інші компоненти. Формулювання задачі через маскування забезпечує вибіркове виділення спектральних областей, пов'язаних з конкретними джерелами, та сприяє впровадженню гібридної схеми обробки, яка поєднує елементи класифікації та регресії в межах єдиної нейронної архітектури. Методологія дослідження включає проектування архітектури мережі, підготовку вхідних даних на основі спектрограм, навчання моделі на багатокомпонентних аудіоміксах і перевірку якості розділення з використанням критеріїв узгодженості реконструкції. Особливу увагу приділено забезпеченню стабільної збіжності моделі та збереженню змістовних акустичних структур у передбачених масках. Отримані результати демонструють стабільну ізоляцію звукових компонентів і сталу ефективність на тренувальному та валідаційному наборах даних. Кількісна оцінка показує точність розділення 0.772 для вокалу, 0.766 для ударних, 0.944 для басів та 0.764 для інших джерел, при цьому відповідні значення середньоквадратичної помилки знаходяться в діапазоні від 0.044 до 0.203 для досліджених категорій. Найвищу ефективність отримано для розділення басів, що пояснюється чітко вираженою низькочастотною спектральною структурою цього джерела. Оцінювання на рівні сигналу за метриками SI-SDR, SDR та SNR показало значення в діапазоні від -1.24 до 4.10 дБ (SI-SDR), від -0.26 до 4.59 дБ (SDR) та від 1.16 до 5.09 дБ (SNR), при цьому найвищі значення спостерігалися для басових і вокальних компонентів, що узгоджується з результатами оцінювання за точністю. Результати підтверджують ефективність поєднання бінарного маскування зі згортковою обробкою спектрограм для обчислювально ефективного розділення звукових джерел. Запропонований підхід, реалізований на основі компактної нейронної архітектури з 323,233 параметрами моделі, може бути застосований у системах музичного виробництва, рішеннях для покращення мовлення, інтелектуальних платформах аналізу аудіоданих та інших середовищах обробки звуку, що потребують надійних і легковагових механізмів розділення.

Ключові слова: комп'ютерні науки, штучний інтелект, згорткові нейронні мережі, аналіз аудіоданих, обробка аудіосигналів, розділення звукових джерел.