Research Archive of
Rising Scholars (preprint)
Where bright minds share their learnings

Predicting the moisture content of a drug mixture based on acoustic signals during the granulation process using deep machine learning

Olzhas Myrzakhmet [1], Ayazbek Adilzhan [2] , Zhaina Mukhametbay [3], Talgat Miras [4]

[1] National School of Physics and Mathematics
[2] National School of Physics and Mathematics
[3] National School of Physics and Mathematics
[4] National School of Physics and Mathematics

Abstract

This study aims to present an acoustic method for real-time monitoring of the pharmaceutical wet granulation process. Granulation is a widely used process in pharmaceutical production, and the quality of the granule mass is directly related to the material and process parameters. The aim of the study is to accurately determine the moisture content of the granule mass and the granulation phases using machine learning tools by analyzing the sound signals recorded by the microphone. According to the hypothesis, the acoustic emissions generated during granulation are sensitive to phase changes, and their analysis by deep neural networks allows us to classify each granulation phase with high accuracy. The study consisted of three main stages. In the first stage, acoustic microphones were used to record sound signals from each granulation phase. In the second stage, the obtained sound data was spectrally transformed and prepared as input to a convolutional neural network. In the third stage, the model was trained and tested, and the accuracy of granulation phase classification was assessed. The novelty of the study is that it demonstrates an effective method for completely non-contact and real-time monitoring of the granulation process. Data collection, pre processing, and model building were performed completely independently. The results showed that up to 94 percent classification accuracy was achieved using microphone data. This acoustic method is a reliable tool for process quality control in pharmaceutical production. The proposed approach meets GMP requirements and allows the development of real-time control systems in automated production, improving product quality.

Keywords: Deep Learning , Real-Time Monitoring , Phase Classification , Convolutional Neural Networks , Wet Granulation , Acoustic Signals , Mel-Spectrograms.

1. Introduction

The production of pharmaceutical products is strictly controlled because substandard products can harm human health. The production process and individual operations are influenced by many factors related to both the process itself and the materials used. Because of this, predicting the exact "endpoint" of the process is difficult; the complexity is increased by the fact that regulations allow for minor variations in the established process. Even with strict production control, it is impossible to completely eliminate quality variability due to differences in raw materials. Therefore, there is a high need for process monitoring and gaining a deeper understanding of what is happening in order to produce a consistent product from batch to batch [1,2].

Currently, various analytical methods are being studied that allow for real-time process monitoring and obtaining more accurate information about what is happening. One such method is

acoustic emission (AE); AE consists of transient acoustic signals that typically occur at frequencies above human hearing. The advantage of AE is that it is a non-invasive, real-time method. Just as optical microscopy [3] or near-infrared spectroscopy [4] capture the electromagnetic signatures of processes, acoustic emission captures the mechanical signatures of occurring events. At the same time, AE sensors do not require direct contact or visibility of the material, and there is no need to make modifications to the production equipment.

It should be noted that traditional in-line sensors penetrating the process working area can influence the process itself, making it impossible to install or upgrade them after the technology has been approved. In this regard, a promising approach within the PAT (Process Analytical Technology) concept is the use of acoustic emission recorded contactlessly using microphones, which can be easily integrated into the production process without interfering with the equipment's working area.

This study considers the high-shear wet granulation process. As the literature shows, acoustic signals generated during granulation can contain important information about material properties and the process itself [5]. Previously, researchers have attempted to correlate these signals with the physical characteristics of the granulated mass, such as mixture density, liquid content, particle size, and compression behavior. Studies have shown that acoustic signals can be used to determine the stages of granulation and the end of the process [6, 7]. For example, ultrasonic signals helped describe the flow in a fluidized bed [8], while sound signals from the acoustic range made it possible to distinguish the stages of high-shear granulation [7,8,9]. The average frequency of the acoustic signal could indicate the end of the process [7], and the power spectral density could indicate critical quality parameters, such as over-wetting [11].

However, existing studies were relatively old and used subjective criteria selection that was only suitable for a specific setup [10]. The application of machine learning methods allows for improved classification accuracy and the elimination of subjectivity.

In our study, a supervised learning method is used to classify the stages of high-shear wet granulation based on acoustic signals [13,14]. Classification is performed depending on the quality of the granulated mass, which is directly determined by the amount of liquid and the compression density. To date, the application of machine learning for monitoring granulation and predicting its endpoint using acoustic emission has barely been studied in the pharmaceutical industry [12].

2. Methodology

This section details the experimental setup, materials used, data collection methods, and the machine learning approach.

2.1 Experimental Setup

The wet granulation procedure was carried out on a laboratory setup. To record acoustic signals, a condenser microphone was used, positioned in front of the mixer bowl. The layout of the setup is shown in Figure 1.

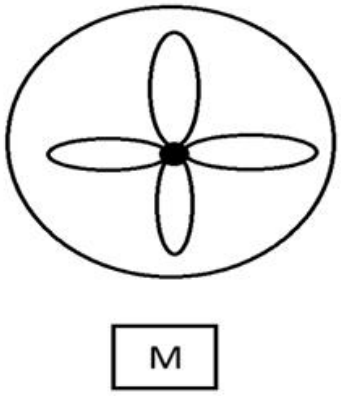

Figure 1. Top view of the granulator and microphone placement. M — condenser microphone.

A Thermomix® TM6 (Vorwerk, Wuppertal, Germany) was used for granulation. To approximate the operation of the laboratory mixer to an intensive industrial one, the original mixing blade was replaced with a specially manufactured blade that ensures high-quality material mixing.
A peristaltic pump was used for the dosed water supply, and the amount of liquid was controlled using a laboratory scale. The experiments were conducted in a quiet laboratory without air conditioning.

2.2. Granulation Process

A binary placebo mixture was used in the experiments: lactose monohydrate and microcrystalline cellulose in an 80:20 ratio. The total mass of the dry mixture was 150 g. Water was supplied at a rate of 2 ml/min. The model formula in Table 1 represents a representative industrial formula.

Material	**Function**
Lactose monohydrate	**Filler**
Microcrystalline cellulose	**Auxiliary component**

Table 1. Model formulation consisting of excipients.

Each of the 6 granulation runs lasted 75 minutes and was divided into three phases based on the moisture level of the granules:

- Dry phase (dry): the first 25 minutes, water was added until reaching 33% moisture content by dry weight (50 g of water).
- Optimal phase (opt): the next 25 minutes of mixing without adding water, the moisture content was maintained at 33%.
- Wet phase (wet): water was added for the first 3.5 minutes (7 g of water) to reach 38% moisture content by dry weight, then mixing continued for another 25 minutes.

Equal-duration intervals ensured a sufficient volume of data and its even distribution across phases. Table 2 summarizes the information about the phases and added water.

Phase	Time [min]	Moisture, ω by dry weight	Notes
dry	**0–25**	**up to 33%**	**water was added continuously**
opt	**25–50**	**33%**	**water was not added**
wet	**50–75**	**up to 38%**	**water was added the first 3.5 min**

Table 2. Division of the granulation process into phases, indicating the time and amount of added water.

2.3. Data Collection

Acoustic data were recorded using a Rode NT-USB condenser microphone placed directly in front of the working area of the mixing bowl to ensure maximum sensitivity to acoustic changes in the granulation process. The operating principle of a condenser microphone is based on converting membrane vibrations caused by sound waves into an electrical signal, which allows capturing a wide range of frequencies and subtle variations in the acoustic profile that occur when the mixture transitions between different modes and phases of granulation [15].

Signal recording was carried out in the Audacity v3.1.3 software environment with a sampling rate of **44.1 kHz**, providing sufficient resolution for subsequent spectral and temporal analysis. Audio files were saved in FLAC format, which eliminated information loss during storage and data preparation for analysis.

Additionally, Audacity was used for the manual labeling of time segments corresponding to key stages of the granulation process. The presence of such labeling was a prerequisite for preparing the dataset intended for training machine learning algorithms in a supervised learning format.

2.4. Data Preparation and Processing

In the first stage of audio signal processing, we transformed it from the time domain to the time-frequency domain using the Short-Time Fourier Transform (STFT) [16]. For this, the signal was divided into small overlapping fragments (windows) with a size of **2048 points** (n_{fft} = 2048) and a step of **512 points** between windows (hop length = 512). Such a transformation makes it possible to determine which frequencies are present in the signal at any given point in time, which is important for analyzing the acoustic characteristics of the granulation process.

To transform the audio signal x[n] into a time-frequency representation, a windowed Fourier transform (STFT) was used, which is calculated using formula (1):

$$X(m,k) = \sum_{n=0}^{N-1} x[n+mH]w[n]e^{-j\frac{2\pi kn}{N}}$$

where:

- X(m, k) is the spectrogram matrix (time m, frequency k);
- x[n] is the input discrete signal;
- w[n] is the Hann window function of length N (**2048 points**);
- H is the window step (hop length, **512 points**).

In the next stage, the obtained spectra were converted to the Mel scale, which approximates the frequency representation to the characteristics of human hearing perception [17]. The Mel scale stretches low frequencies and compresses high ones, making sound features more informative for subsequent machine learning. For the spectrogram, **128 Mel coefficients** were selected, each of which corresponds to the sound intensity in a specific frequency band [17-19].
The conversion of linear frequency f (Hz) into the Mel scale (Mel), taking into account the psychoacoustic features of perception, was performed according to formula (2):

$$f_{mel} = 2595 \log_{10}(1 + \frac{f}{100})$$

After that, the Mel-spectrograms were divided into non-overlapping windows of **32 time columns** to form manageable data fragments for the neural network. Each window was log-transformed, which allowed bringing the spectrum amplitude to a logarithmic scale close to human loudness perception, while simultaneously reducing the spread of values for more stable model training. As a result, each spectrogram fragment had a size of **128 × 32** and was treated as a separate sample for input into the neural network. For all stages of data processing, the librosa library [20] in Python was used, which provides convenient STFT calculation, Mel-spectrogram construction, logarithmization, and window splitting.
This approach allows the neural network to analyze short, standardized fragments of the acoustic signal and extract features necessary for classifying granulation phases or other acoustic events in the process.

25. Neural Network Architecture (CNN)

To automatically extract hierarchical features from logarithmic Mel-spectrograms, a deep architecture based on the VGG (Visual Geometry Group) topology was developed [21]. This approach eliminates the stage of manual hand-crafted features, delegating the task of finding relevant acoustic patterns to the backpropagation algorithm.

2.5.1. Network Topology

The network architecture is a cascade of four sequential Convolutional Blocks, followed by a classification perceptron (Fully Connected layers). The general scheme of the model is shown in Table 3.
The input tensor has dimensions 128 **×** 32 **×** 1 (frequency bins **×** time frames **×** channels).

1. Convolutional Blocks (Feature Extractor):
Each of the four blocks consists of two sequential convolutional layers with a small receptive field (3**×**3). The use of a cascade of small filters makes it possible to increase the effective receptive field of the network and add nonlinearity while maintaining computational efficiency.

Mathematically, the convolution operation for the input feature map X and kernel W is described by formula (4):

$$Y(i,j) = (X * W)(i,j) + b = \sum_m \sum_n X(i+m, j+n) * W(m,n) + b$$

where b is the bias. All convolutions were performed with stride = 1 and zero-padding to preserve spatial dimensions.

2. Nonlinearity and Normalization:
After each convolution operation, Batch Normalization (BN) was applied to stabilize the distribution of activations and accelerate convergence. A Rectified Linear Unit (ReLU) was used as the activation function, defined by formula (5):

$$f(x) = \max(0, x)$$

3. Dimensionality Reduction (Pooling):
At the end of each block, a Max-Pooling subsampling operation with a 2×2 window was applied, which reduced the spatial dimensions of the feature maps by a factor of 4, providing the model with invariance to small temporal and frequency shifts.

4. Classifier:
Deep features were aggregated through high-capacity Dense layers (1024 neurons each). To prevent overfitting in these layers, Dropout regularization with a coefficient p=0.5 (randomly dropping 50% of connections at each training iteration) was applied.

Блок / Слой	Конфигурация слоев и параметры	Выходная размерность
Input	—	128 × 32 × 1
Conv Block 1	2 × [Conv2D (3 × 3, 64 filters) + BN + ReLU] MaxPooling (2 × 2)	64 × 16 × 64
Conv Block 2	2 × [Conv2D (3 × 3, 128 filters) + BN + ReLU] MaxPooling (2 × 2)	32 × 8 × 128
Conv Block 3	2 × [Conv2D (3 × 3, 256 filters) + BN + ReLU] MaxPooling (2 × 2)	16 × 4 × 256
Conv Block 4	2 × [Conv2D (3 × 3, 512 filters) + BN + ReLU] MaxPooling (2 × 2)	8 × 2 × 512
Classifier	Flatten / Global Average Pooling Dense (1024 units, ReLU) + Dropout (0.5)	1024
	Dense (1024 units, ReLU) + Dropout (0.5)	
Output	Dense (3 units) + Softmax	3 (Вероятности классов)

Table 3. Detailed specification of the VGG-style CNN architecture.

2.5.2. Probabilistic Inference

The output layer consists of 3 neurons corresponding to the Dry, Opt, and Wet phases.

The transformation of the neural network outputs z_i into a probability distribution P(y = i) is carried out by the Softmax function (6):

$$P(y = i|x) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

where K=3 is the number of classes.

2.6. Training Setup

The model was trained using the stochastic gradient descent method with the adaptive Adam optimizer [22]. The optimizer hyperparameters were set according to recommendations for audio tasks:

- Momentum coefficients: $:\beta_1 = 0.9,\ \beta_2 = 0.999.$
- Learning Rate: 10^{-3}
- Batch size: **128**.

Cross-Entropy Loss was minimized as the objective function. To prevent overfitting, an Early Stopping strategy was used: the training process was interrupted if the accuracy on the validation set did not increase for 20 consecutive epochs.

The implementation was performed in the Python environment using the TensorFlow [23] and Keras deep learning libraries.

3. Results and Discussion

Below are the results of the process state recognition. We sequentially consider the characteristics of the experimental runs, the applied data splitting strategies, and the final classification accuracy achieved by the CNN models for the microphone data.

3.1. Granulation Results

In total, we conducted 6 granulation experiments. Looking at the obtained granules (Figure 2), we noticed a clear pattern: in the very first runs, the granules turned out to be noticeably larger than in the subsequent ones.

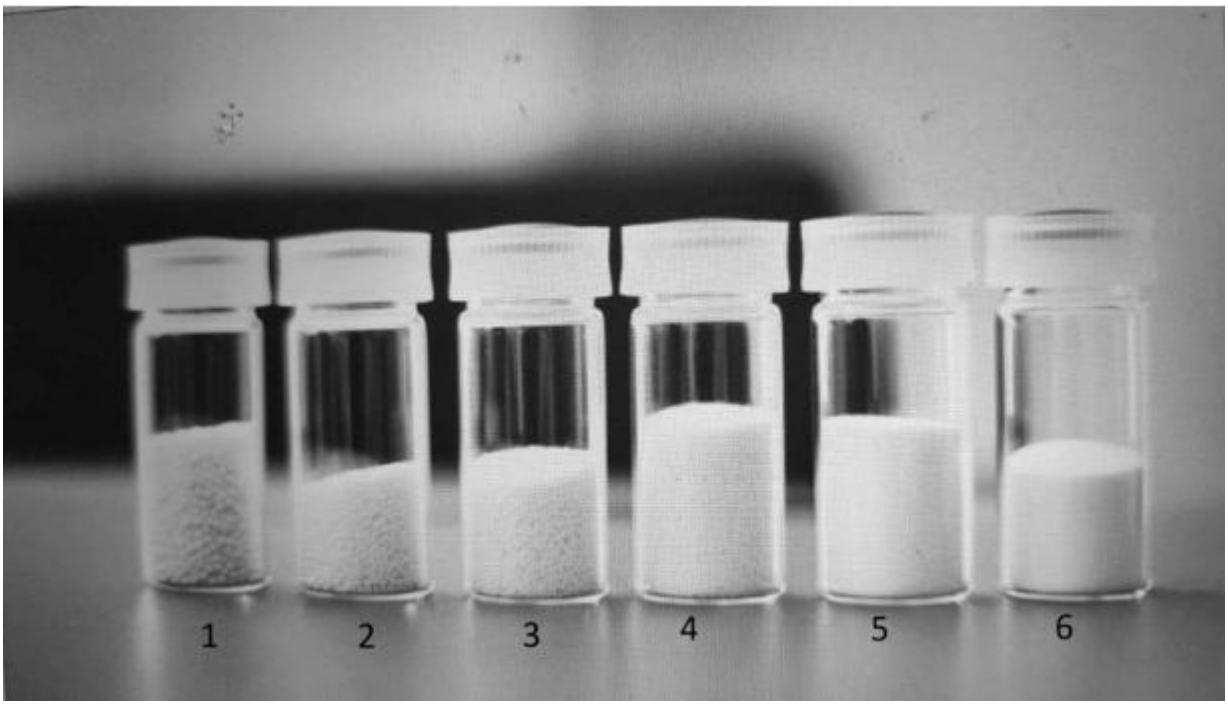

Figure 2.

We attribute this to the fact that at the beginning of the operation, the system is still cold, and the mixing process proceeds slightly differently due to the heating of the equipment. In practice, excessively large granules are considered undesirable, as this is a sign of a suboptimal process. Since our goal is to teach the neural network to recognize the correct process, we decided to filter the data. We selected for analysis only those experiments where the granules turned out to be of a normal (small) size. Thus, the first "unstable" runs were excluded, and for training, we used a clean dataset from the remaining stable runs.

3.2. Training and Testing Scenarios

Based on the selected "good" recordings, we devised three testing schemes for our neural network to see how it performs under different conditions:

- Maximum Coverage (Multi-container): In this scenario, we combined the data from most of the experiments: recordings from containers No. 2, 3, 4, and 5 were used to train the neural network, while the recording from container No. 6 was left for the final test. This approach implements the principle of "the more data, the better," providing the model with a maximum of diverse information for adjusting the weights.
- Group Split (Batch Split): Here, we applied a more rigorous approach to testing. Recordings from one group of batches (containers No. 4, 5, 6) were used for training, while testing was conducted on a completely different group (containers No. 2, 3) that the model had not previously "seen." This is the most realistic scenario for production, showing whether a system tuned on one set of batches can work correctly with others.
- Minimal Data (Little Data): To test the sensitivity of the method to the volume of data, we trained the model using the recording of only one experiment. Testing was conducted on all other available recordings. This study helps to understand: is it necessary to spend resources on collecting a large database, or is one short example (75 minutes) enough for the neural network to confidently grasp the difference between the dry and wet phases.

3.3. Overall Classification Accuracy

To evaluate the performance of the developed convolutional neural network (CNN), the Classification Accuracy metric was used. It is calculated as the ratio of the number of correctly recognized samples to the total volume of the test set.

Table 4 presents the final results for the three experimental scenarios described in Section 4.2. Since a single data source—a condenser microphone—was used in this study, the comparison is made exclusively between the training strategies.

Training Scenario	**Accuracy [%]**
1. Maximum Coverage (Multi-container)	**94.2**
2. Group Split (Batch Split)	**93.1**

Training Scenario	Accuracy [%]
3. Minimal Data (Little Data)	**62.7**

Table 4. CNN testing accuracy for various training scenarios (Condenser microphone).

Analysis of the results:

1. Maximum coverage and group split showed consistently high results (above 93%). The minimal difference between them (1.1%) confirms the excellent generalizing ability of the model: it successfully works with new batches that were not involved in the training.
2. However, in the minimal data mode, the accuracy dropped sharply to 62.7%. This indicates the variability of acoustic signals between runs. The recording of a single experiment proved to be insufficient for the neural network to learn robust phase features.

3.4. Temporal Dynamics of Phase Classification

This section provides a detailed analysis of the probability output of the convolutional neural networks, allowing for an assessment of the model's confidence degree when classifying samples in the dynamics of the process.

Figure 3 visualizes the performance of the network trained on the condenser microphone data under the "Maximum Coverage" (Multi-container) scenario. The curves display the change in the probability of the process belonging to a specific rheological phase over time:

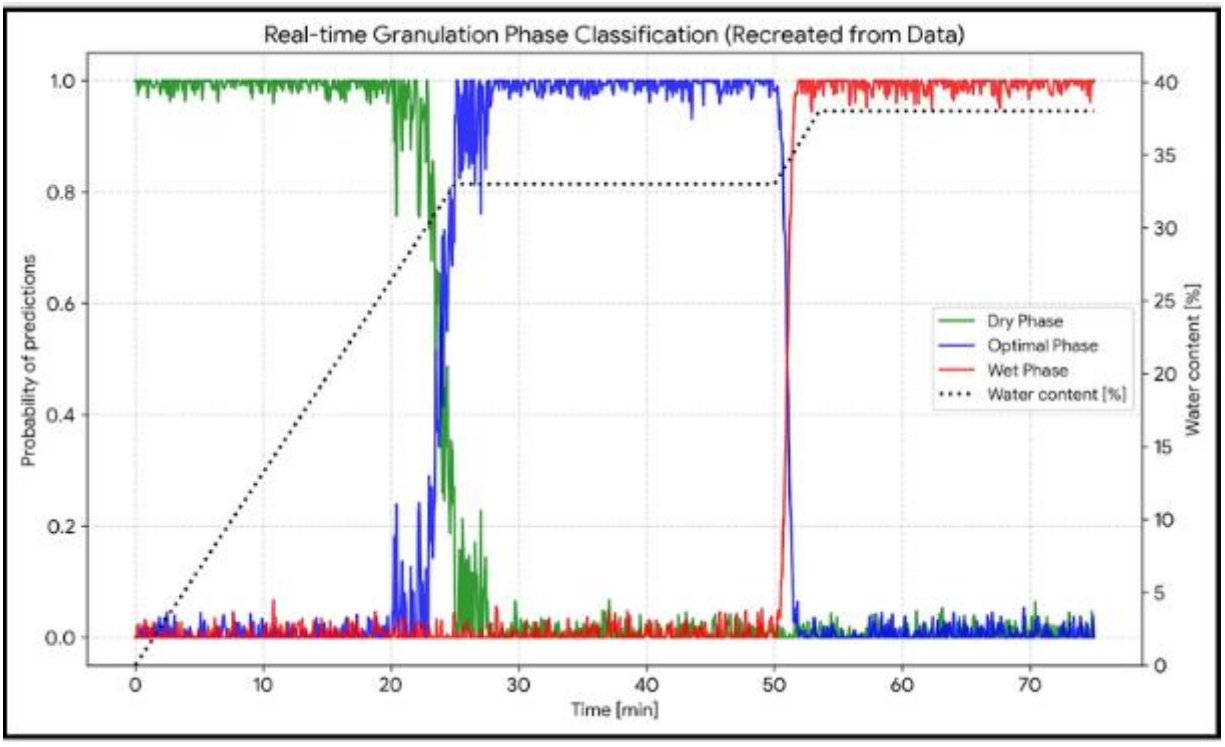

Figure 3. Change in the probability of classifying process states over time (condenser microphone data, "Maximum Coverage" scenario).

The solid colored lines show the network's confidence in the current class: **blue** — Dry phase, **green** — Opt phase, **red** — Wet phase. The dashed black line displays the actual moisture content in the mixture. It can be seen that the network clearly responds to phase transitions (at the 25th and 50th minutes), correlating with the change in moisture.

The graph (Figure 3) shows that the network correctly recognizes the current phase at all stages:

1. **Beginning (0–22 min):** In the first 20 minutes, the network gives the highest probability to the Dry state (blue line). The probability that the mixture is wet (Wet) is practically zero here. Sometimes the network slightly "doubts" and gives a small chance to the Optimal state, but the primary choice remains correct.
2. **First transition (22–27 min):** Here we see a clear change of the leader: the blue line goes down, and the green one (Optimal) goes up. It is noticeable that this transition does not happen instantly and is a bit "nervous" (the lines jump) — this means that the wetting process is uneven, and the network is trying to catch the moment of change.
3. **Middle (up to 50 min):** After the transition and up to the 50th minute, the network behaves very stably. It is confident that the process is in the Optimal state, and practically does not consider other options.
4. **Second transition (after 50 min):** A sharp change occurs here. The probability of the optimal phase quickly drops to zero, and the red line (Wet) shoots up just as quickly. This transition looks much smoother and clearer compared to the first one. Until the very end of the experiment, the network continues to confidently identify the mixture as wet.

Summary: Our neural network successfully copes with the task. It does not just guess the phase on average, but also accurately shows the time boundaries of each stage, as well as the nature of the transition from one state to another.

4. Conclusion

In this work, we explored how machine learning helps distinguish granulation phases based solely on the sound recorded by a conventional condenser microphone. We fed sound spectrograms to the input of a neural network (CNN) and set the goal of teaching the system to automatically determine the quality of the process.

We managed to achieve reliable phase prediction in all scenarios where multiple recordings were used for training. In such cases, the classification accuracy consistently exceeded **90%**. However, when we restricted the training to data from just one experiment, the accuracy decreased. This confirms a well-known rule of machine learning: for a model to work well, the data must be as diverse as possible.

It is important to note the advantage of our method: the neural network analyzes the entire sound spectrum as a whole, which saves the operator from having to manually search for the "right" frequencies or adjust thresholds.

Important tasks remain outside the scope of this study. In the future, it is necessary to check whether this method will work in the noisy conditions of real pharmaceutical production, as well as to find out whether this model can be used for quality control with minor changes in the mixture's composition. Finally, a promising goal is the refinement of the algorithm for real-time operation, so that the system itself can signal the end of the process.

References

1. **S. Ramm, R. Fulek, V. A. Eberle, C. Kiera, U. Odefey, and M. Pein-Hackelbusch**, "Compression density as an alternative to identify an optimal moisture content for high shear wet granulation as an initial step for spheronisation," *Pharmaceutics*, vol. 14, no. 2303, 2022, doi: 10.3390/pharmaceutics14102303.
2. **T. Reimers, J. Thies, P. Stöckel, S. Dietrich, M. Pein-Hackelbusch, and J. Quodbach**, "Implementation of real-time and in-line feedback control for a fluid bed granulation process," *Int. J. Pharm.*, vol. 567, p. 118452, Aug. 2019, doi: 10.1016/j.ijpharm.2019.118452.
3. **R. Attota, P. P. Kavuri, H. Kang, R. Kasica, and L. Chen**, "Nanoparticle size determination using optical microscopes," *Appl. Phys. Lett.*, vol. 105, p. 163105, Oct. 2014, doi: 10.1063/1.4900484.
4. **M. Jamrógiewicz**, "Application of the near-infrared spectroscopy in the pharmaceutical technology," *J. Pharm. Biomed. Anal.*, vol. 66, pp. 1–10, Jul. 2012, doi: 10.1016/j.jpba.2012.03.009.
5. **B. Liu, J. Wang, J. Zeng, L. Zhao, Y. Wang, Y. Feng, and R. Du**, "A review of high shear wet granulation for better process understanding, control and product development," *Powder Technol.*, vol. 381, pp. 204–223, Mar. 2021, doi: 10.1016/j.powtec.2020.11.051.
6. **M. Whitaker, G. Baker, J. Westrup, P. Goulding, R. Belchamber, and M. Collins**, "Applications of acoustic emission to the monitoring and end point determination of a high shear granulation process," *Int. J. Pharm.*, vol. 205, pp. 79–91, 2000, doi: 10.1016/S0378-5173(00)00479-8.
7. **L. Briens, D. Daniher, and A. Tallevi**, "Monitoring high-shear granulation using sound and vibration measurements," *Int. J. Pharm.*, vol. 331, pp. 54–60, 2007, doi: 10.1016/j.ijpharm.2006.09.012.
8. **H. Tsujimoto, T. Yokoyama, C. Huang, and I. Sekiguchi**, "Monitoring particle fluidization in a fluidized bed granulator with an acoustic emission sensor," *Powder Technol.*, vol. 113, pp. 88–96, 2000, doi: 10.1016/S0032-5910(00)00205-9.
9. **D. Daniher, L. Briens, and A. Tallevi**, "End-point detection in high-shear granulation using sound and vibration signal analysis," *Powder Technol.*, vol. 181, pp. 130–136, 2008, doi: 10.1016/j.powtec.2006.12.003.
10. **E. M. Hansuld, L. Briens, J. A. B. McCann, and A. Sayani**, "Audible acoustics in high-shear wet granulation: Application of frequency filtering," *Int. J. Pharm.*, vol. 378, pp. 37–44, 2009, doi: 10.1016/j.ijpharm.2009.05.042.
11. **E. Hansuld, L. Briens, A. Sayani, and J. McCann**, "Monitoring quality attributes for high-shear wet granulation with audible acoustic emissions," *Powder Technol.*, vol. 215–216, pp. 117–123, 2012, doi: 10.1016/j.powtec.2011.09.034.
12. **H. Lou, B. Lian, and M. J. Hageman**, "Applications of machine learning in solid oral dosage form development," *J. Pharm. Sci.*, vol. 110, pp. 3150–3165, 2021, doi: 10.1016/j.xphs.2021.04.013.
13. **Y. LeCun, Y. Bengio, and G. Hinton**, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.
14. **J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell**, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017, doi: 10.1109/TPAMI.2016.2599174.

15. **C. N. Teague, S. Hersek, H. Toreyin, M. L. Millard-Stafford, M. L. Jones, G. F. Kogler, M. N. Sawka, and O. T. Inan**, "Novel methods for sensing acoustical emissions from the knee for wearable joint health assessment," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 8, pp. 1581–1590, Aug. 2016, doi: 10.1109/TBME.2016.2543226.
16. **S. W. Smith**, *Digital Signal Processing: A Practical Guide for Engineers and Scientists*, vol. 1. Oxford, UK: Newnes, 2003.
17. **Y. Qu, X. Li, and Z. Qin**, "Acoustic scene classification based on three-dimensional multi-channel feature-correlated deep learning networks," *Sci. Rep.*, vol. 12, p. 13730, 2022, doi: 10.1038/s41598-022-17938-0.
18. **A. Mesaros, T. Heittola, and T. Virtanen**, "Acoustic scene classification: An overview of DCASE 2017 challenge entries," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, 17–20 Sep. 2018, pp. 411–415, doi: 10.1109/IWAENC.2018.8586003.
19. **T. Zhang, G. Feng, J. Liang, and T. An**, "Acoustic scene classification based on Mel spectrogram decomposition and model merging," *Appl. Acoust.*, vol. 182, p. 108258, 2021, doi: 10.1016/j.apacoust.2021.108258.
20. **B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto**, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python in Science Conf.*, Austin, TX, USA, 6–12 Jul. 2015, pp. 18–25.
21. **K. Simonyan and A. Zisserman**, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, May 7–9, 2015.
22. **C. R. Harris et al.**, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.
23. **M. Abadi et al.**, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems,” 2015. [Online]. Available: https://www.tensorflow.org