Deep convolutional neural network for detection of pathological speech

Authors: Lukáš Vavrek, Máté Hireš, Dinesh Kumar, Peter Drotár

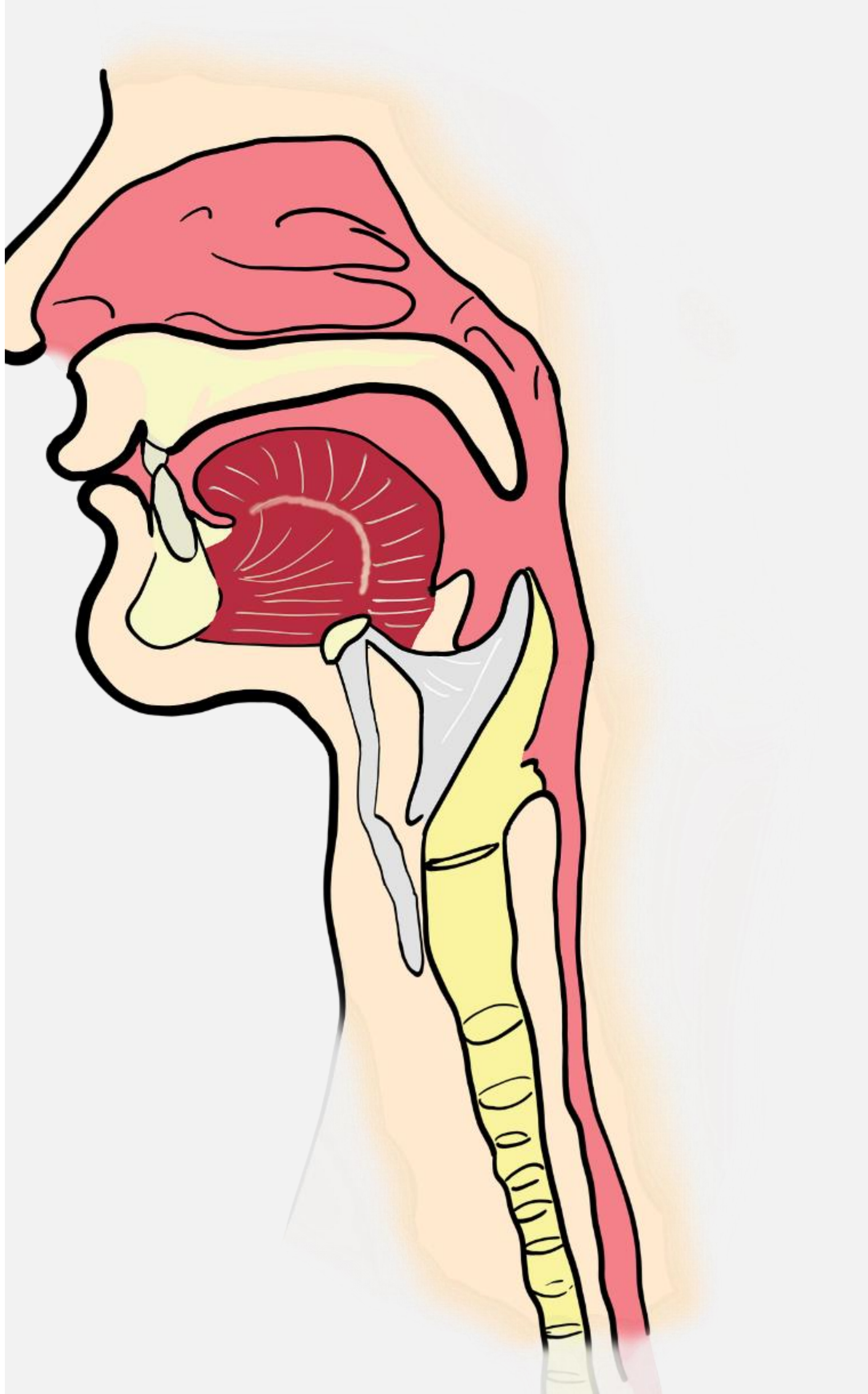
Pathological speech

Voice pathologies

- Affect the ability of larynx to produce voice
- Irregular vocal cord vibrations
- Manifested by changes in voice
- Crucial to diagnose early

Diagnostics

- Required sophisticated medical equipment and trained specialist
- Time-consuming and expensive
- Result is highly dependent on specialist’s experience and skill

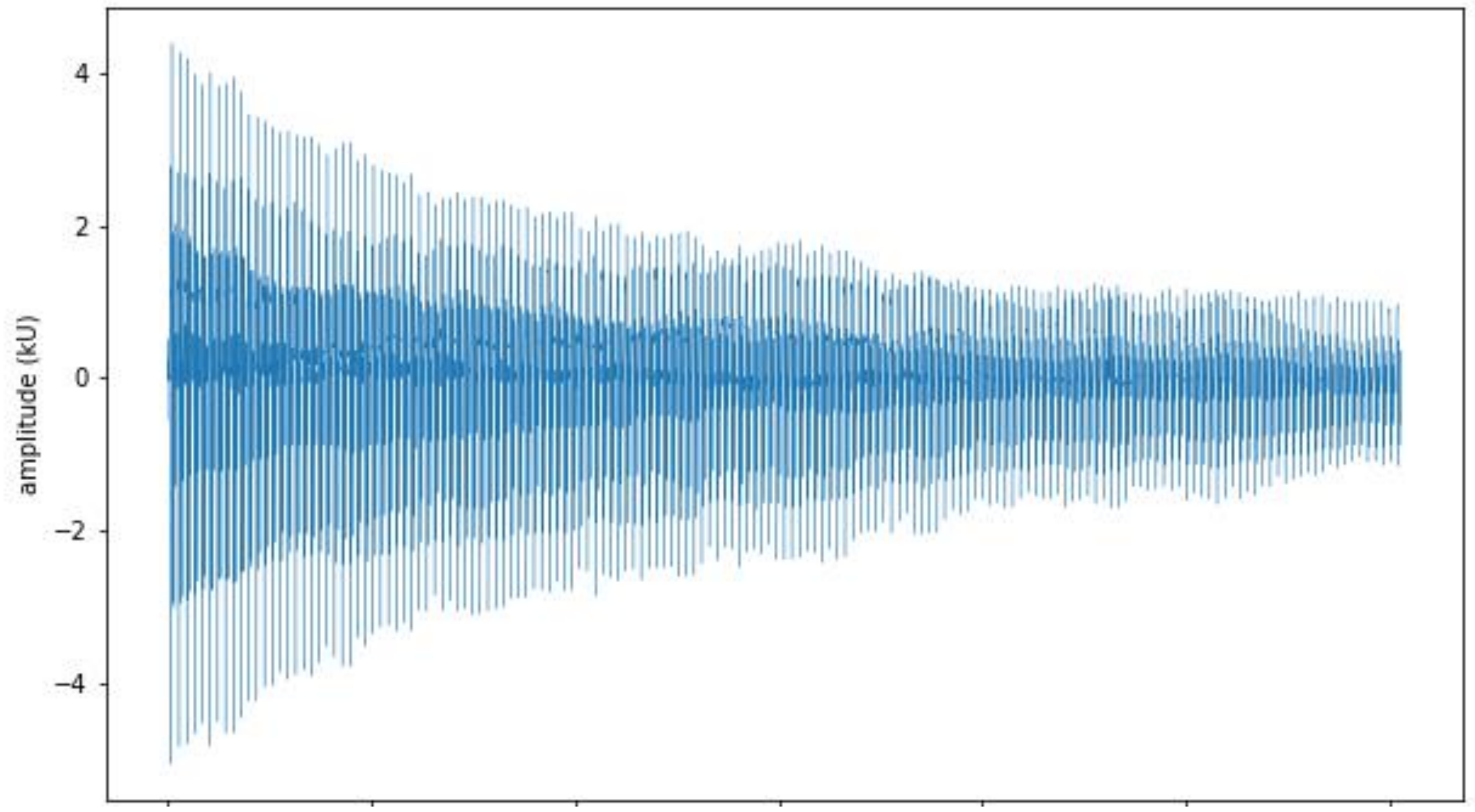


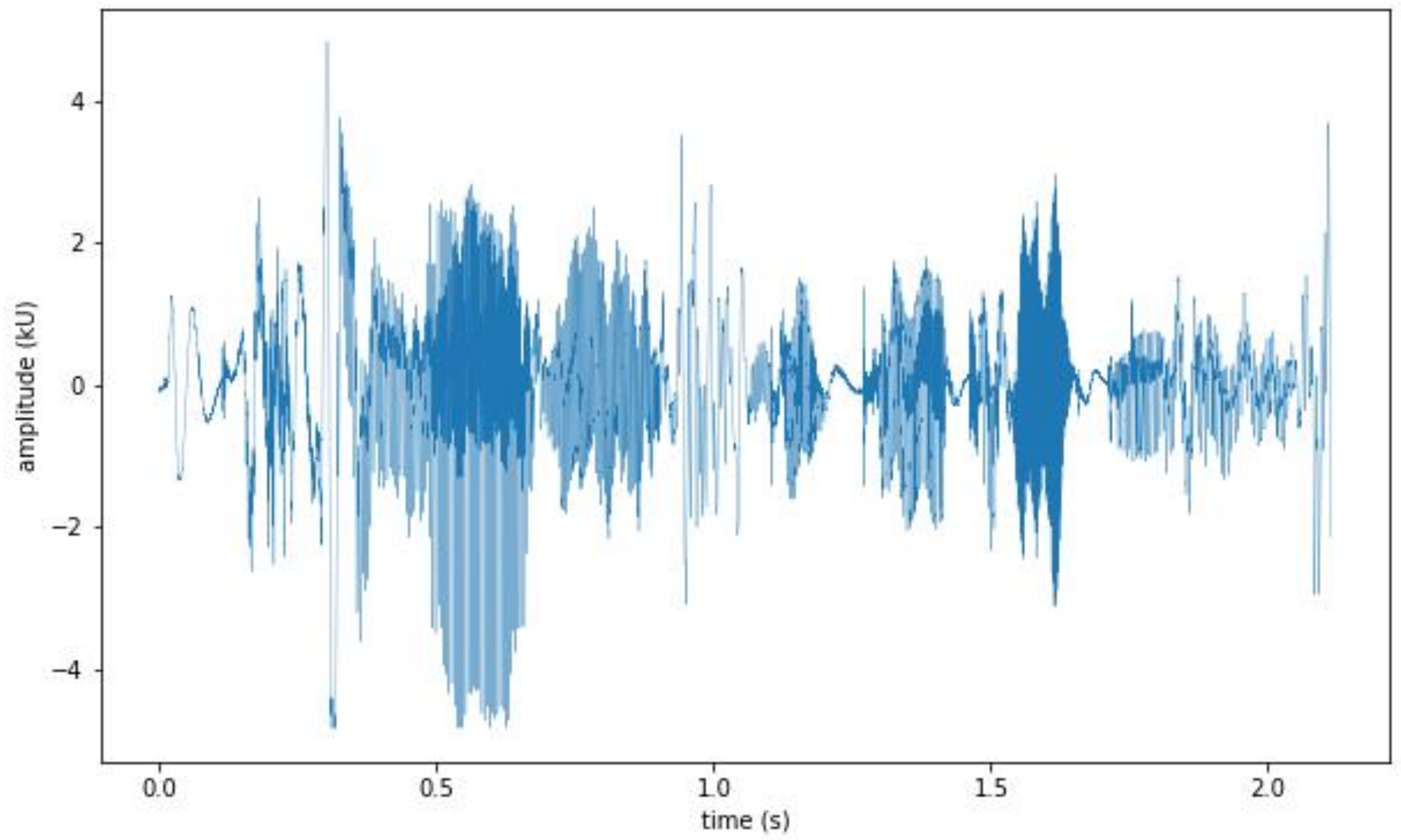
Dataset

Saarbruecken voice database

- Voice recordings from more than 2000 person
- Recordings of vowels /a/, /i/, /u/ produced in normal, low, high and rising-falling pitch, and a sentence: “Guten Morgen, wie geht es Ihnen?“

- Reduced dataset:
    - only organic dysphonia pathologies
    - 506 pathological and 506 healthy subjects

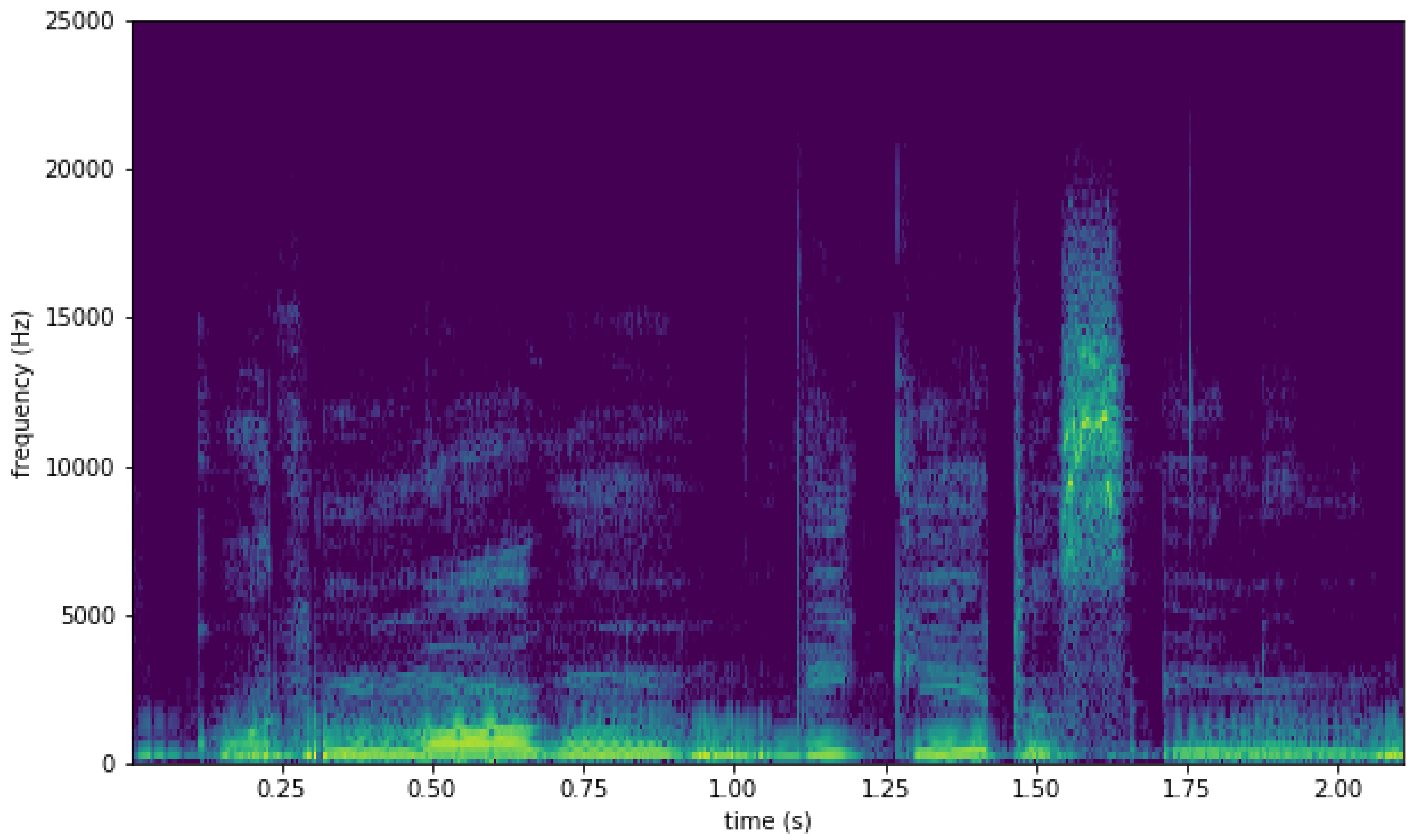




Data preprocessing

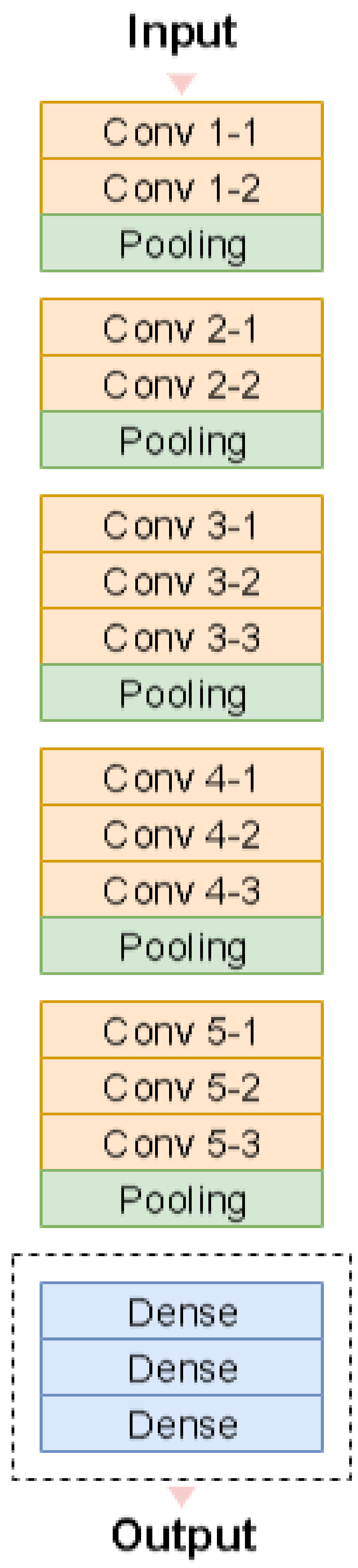
- Conversion to spectrograms using Short-time Fourier transform operation (STFT)
- Visualizing the frequency of the sound over time
- Amplitude is preserved as color intensity

- Data divided into training (60%), validation (20%) and test (20%) sets, using stratified splits
    - Proportion of values in produced groups stays consistent according to provided data



Transfer learning approach

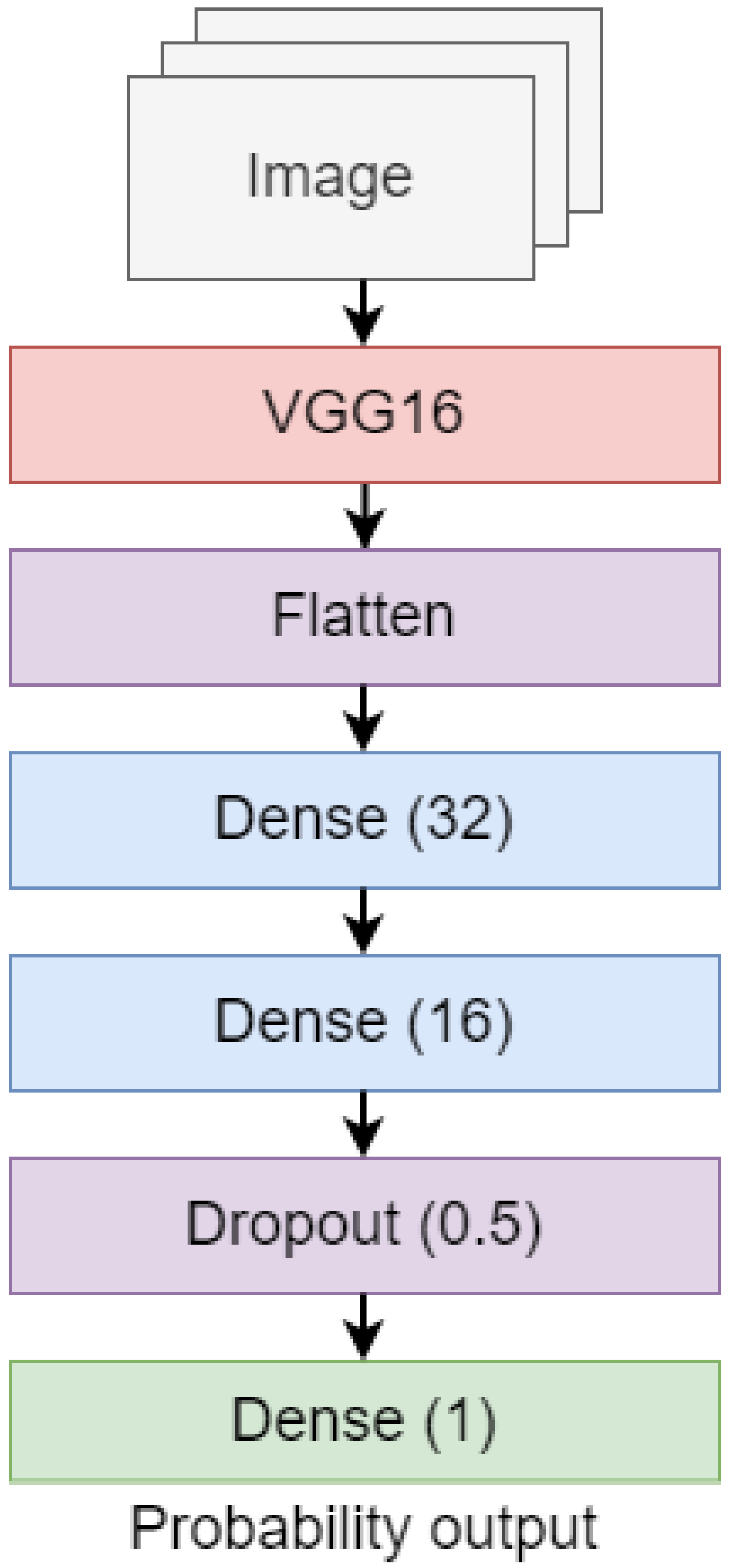
- Reusing existing knowledge of a pre-trained network
- **VGG16 CNN** base model
  - Deep convolutional neural network for object recognition
  - Pre-trained on ImageNet dataset
- Top layers are removed from the base pre-trained network loaded with weights
- **Layers** in base network **are frozen** (won’t be updated)
- Custom classifier on top of base network



CNN single vowel approach

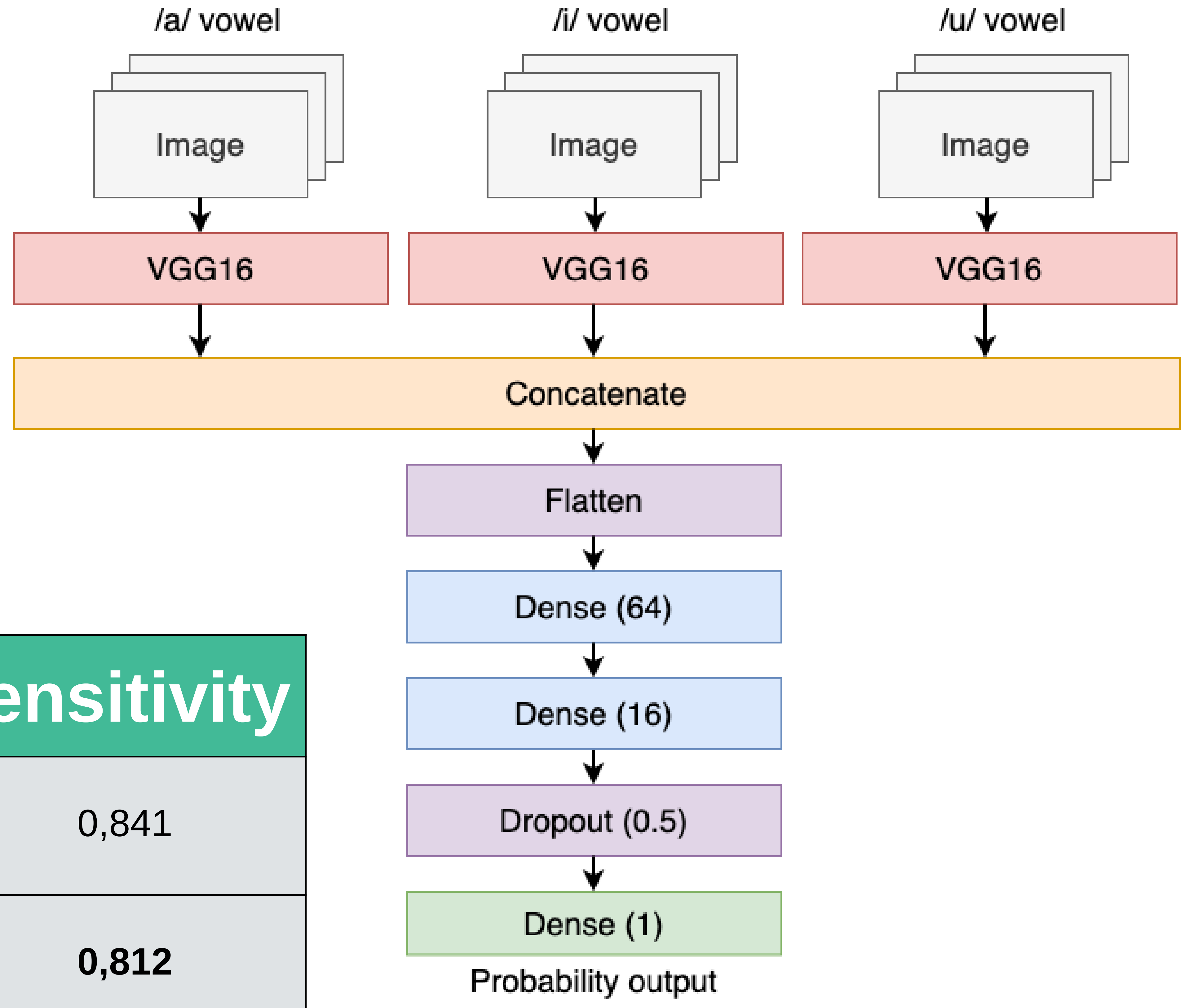
- Straightforward and simple solution
- VGG16 base network with custom classifier
- /a/ vowel data subset of natural modulation
- Only small subset of available data is utilized

	Accuracy	Specificity	Sensitivity
Single layer classifier	74,23%	0,763	0,723
Enhanced classifier with two layers	**79,14%**	**0,775**	**0,807**



Multi-input model with one CNN per input

- Network expects three inputs for each subject
- Each input is processed separately from other inputs
- Results from VGG16 networks are concatenated within the model graph
- Massive network – slow training

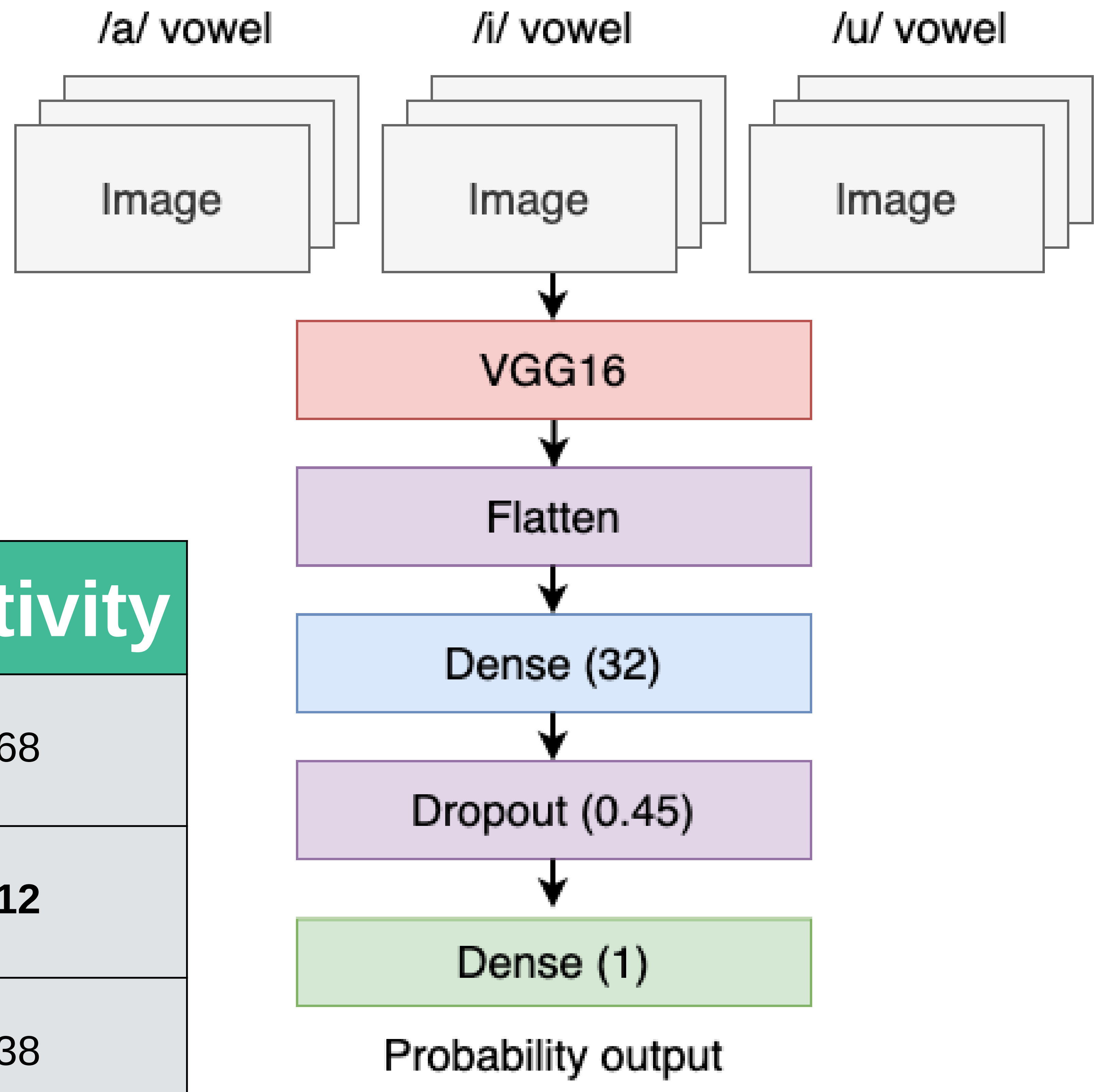


	Accuracy	Specificity	Sensitivity
Multi-input model, two dense layers with 32 and 16 neurons	74,8%	0,657	0,841
Multi-input model, two dense layers with 64 and 16 neurons	**76,3%**	**0,714**	**0,812**

Encoding multiple inputs into image channels

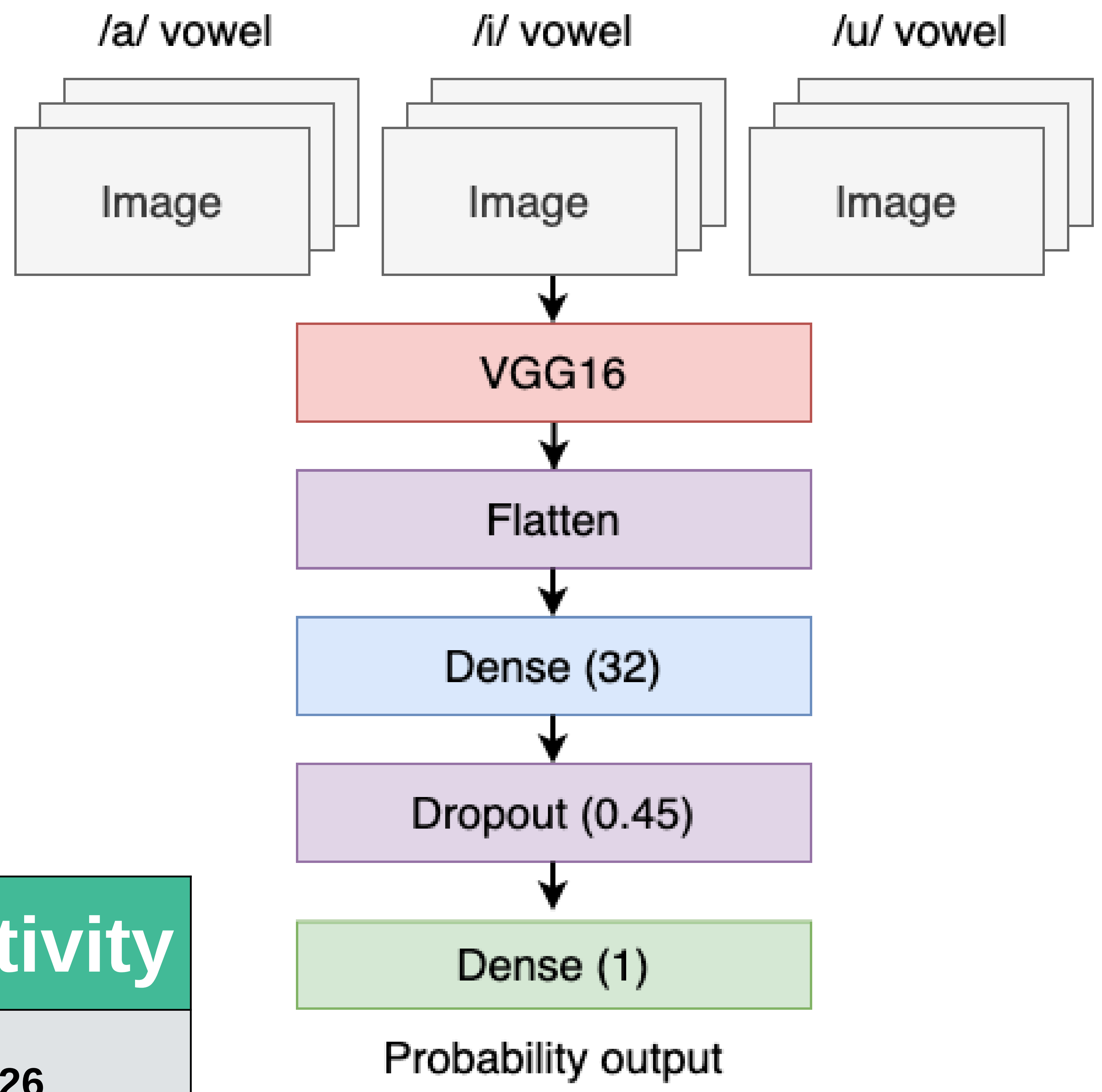
- Combination of two initial experiments
- Small network, with three vowel subset
- Spectrograms of each vowel are combined into a single image, using RGB channels to hold them
- Network was unable to use full potential of data

	Accuracy	Specificity	Sensitivity
Encoded multiple inputs, single dense layer with 128 neurons	72,67%	0,686	0,768
Encoded multiple inputs, single dense layer with 32 neurons	**75%**	**0,686**	**0,812**
Encoded multiple inputs, single dense layer with 32 and 16 neurons	73,38%	0,829	0,638



Fine tuning model with multiple inputs

- Explores effects of fine-tuning of pre-trained network
- Layers are unfreezed from the end of the base CNN
- Weights of unfreezed layers are adjusted during a weight update procedure
- Model can better adapt to a destination problem
- More than 2% increase in accuracy



	Accuracy	Specificity	Sensitivity
Fine tuned model (last three layers)	76,98%	0,714	0,826

Model ensemble

- Combines an advantage of using a bigger data subset with all vowels and using multiple simple models
- The ensemble is composed of the same networks (from single vowel approach)
- Each model is trained separately on a different data subset
- For final prediction, partial answers are combined using a weighted average method
- A prediction weight is assigned to each model based on its evaluation
- **Accuracy improved by more than 2%, while using four times less data**

	Accuracy	Specificity	Sensitivity
Model ensemble	**82,01%**	**0,843**	**0,797**