# Circadian Rhythm in the Brain

Maxi Fischer[1], Elizaveta Korotkova[2], Kateryna Peikova[2], Katharina Fogelberg[3]

**Abstract**
Circadian rhythm is a biological process following the daily cycle of changes in environment, regulated by the ”master clock” in the suprachiasmatic nucleus. In this report we explore a dataset of microscopic fluorescence images of the SCN of a mouse brain, representing the changes in expression of the circadian clock gene Period1 over time. We investigate the regularity and periodicity of the signal, the groups formed by the cells, the effect of proximity on cell behaviour, the movement of the cells during the experiment, and also try to approach the discovery of circadian rhythm as a decoding task. We apply descriptive statistics, unsupervised and supervised machine learning algorithms. We find that the cells may be organized into several interpretable clusters, which follow slightly different patterns and periods, and that proximity of cells within a cluster is also a factor in similar behaviour. Decoding time from the brain images yields satisfactory results. Finally, we show that the observed cells move only slightly during the experiment.

**Keywords**
Computational Neuroscience — Circadian Rhythm — Suprachiasmatic Nucleus

[1] *Hasso Plattner Institute, University of Potsdam, Germany*
[2] *Institute of Computer Science, University of Tartu, Estonia*
[3] *Institute of Geography, University of Augsburg, Germany*

## Introduction

Most land animals have developed an internal rhythm under daily cyclical changes produced by the earth's rotation, such as daily fluctuations in light and temperature. Because of the daily changes in the environment, the rhythms tune the internal physiology to external conditions [1]. These are called circadian rhythms.

A circadian rhythm is a biological process that displays a repetitive pattern one iteration of which takes roughly 24 hours, meaning that these processes follow a daily cycle. Circadian rhythms are found in many organisms, including not only animals, but also plants and fungi.

The so-called biological clock is an internal "timing device" of an organism. In vertebrates, a formation of about 20,000 neurons called the suprachiasmatic nucleus (SCN), located in the hypothalamus, is believed to perform the role of a "master clock" [2]. It is the main generator of circadian rhythms in mammals [3] and controls the release of melatonin in the epiphysis, e.g. to establish sleeping patterns by enforcing a metabolic rhythm or to synchronize the biological clock in an organism. The activity of the SCN neurons changes periodically during the day and it adjusts to external light signals delivered by the optic chiasm, which is connected to the eye. At the same time, other organs and even cells in an organism have their own circadian clocks.

The aim of this project is to investigate an existing dataset consisting of fluorescence images, which shows the expression of the circadian clock gene Period1 in the suprachiasmatic nucleus of a mouse. Mice also follow the internal clock. That makes the results comparable to the human circadian clock, although mice are nocturnal. This only causes a shift in the circadian clock itself, but should not influence the periodicity [3]. The primary challenge set by the creators of this dataset is the necessity to predominantly use automatic processing to avoid performing a lot of manual analysis.

Within this project we followed several different, but interconnected directions of exploratory analysis. First, we used unsupervised machine learning methods to investigate whether it is possible to discover meaningful clusters within the observed brain region (Section 1). At the next step, using signal processing techniques, we studied how regular the expression in the SCN generally is and what dominant periods and patterns it has in the discovered clusters (Section 2). We also researched how the distance between cells affects the similarity of their fluorescence patterns, focusing again on both the entire image and the single clusters (Section 3). Then, with supervised machine learning algorithms, specifically random forests, we attempted to approach the discovery of a daily cycle as a decoding task, trying to predict time of day using features extracted from brain images (Section 4). Finally, we investigated how single cells can be effectively detected in the images and whether those cells had changed their locations during the seven-day experiment (Section 5).

## Dataset Description

The dataset explored in this report is provided on Kaggle[1] by researchers of the Charles Allen lab at the Oregon Institute of Occupational Health Sciences[2]. It is a time-series of microscopic fluorescent images of the suprachiasmatic nucleus, which is located in the hypothalamus in the brain (Figure 1). A fluorescent protein is used as a reporter for the expression of the gene Period1, which is important to the maintenance of circadian rhythms in cells.

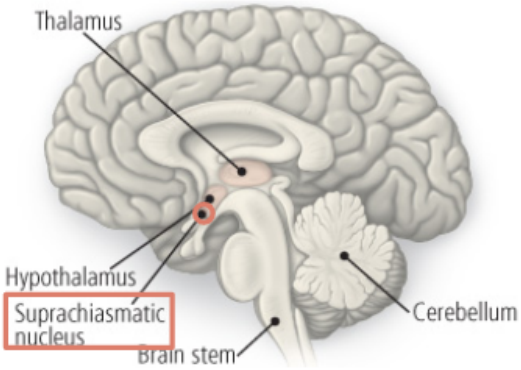



**Figure 1.** Location of the SCN in the brain

The expression of the gene is observed over seven days, to follow the changes through multiple daily cycles. The researchers' goal in gathering this dataset was to gain an understanding of how the network in the suprachiasmatic nucleus synchronizes the individual neurons' clocks to produce a common circadian rhythm. The challenge is to automate the analysis so as not to analyze each image in the time-series manually to follow the experiment.

The dataset consists of four files. One is a CSV file mapping image numbers and frame numbers to how many hours had passed since the beginning of the experiment when the image was taken. Two are in Numpy's [4] compressed array format (NPZ), each of them contains an array of image IDs and an array of $512 \times 512$ image representations with each pixel being represented by its fluorescence intensity (see Figure 2). There is a total of 1685 images, corresponding to ten data points per hour, obtained over 7 days. The fourth file is a compressed ZIP folder of raw fluorescence images.

Several kernels[3] are also provided to accompany the dataset. Our explorations are largely based on code and suggestions from those kernels.

## 1. Clustering

One of the most important tasks is to divide the given pictures into anatomical clusters of the mouse brain. We assume that different parts of the brain behave differently depending on the type of cells and number of cells in that region and their interactions. That might cause different circadian rhythm patterns for each cluster.

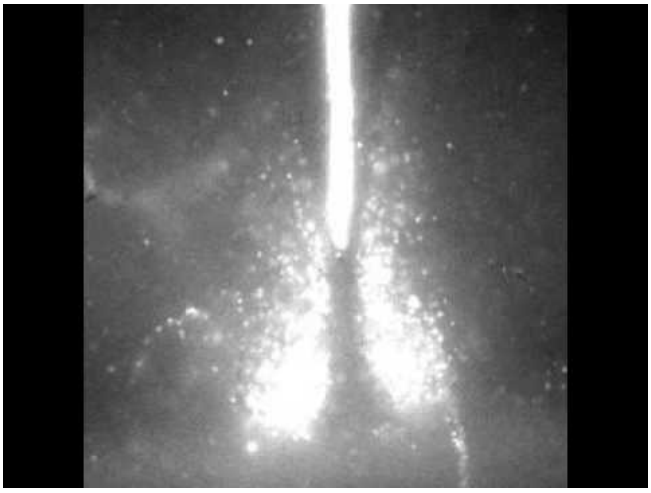


**Figure 2.** An example image from the dataset. The vertical line in the middle of the picture shows the third ventricle. The two bright regions on either side of the ventricle are the two portions of the Suprachiasmatic nucleus (SCN). Below the ventricle and the SCN is a dark horizontal band that represents the optic chasm.

The function of the **suprachiasmatic nucleus** is outlined in detail in the Introduction. It is a region within the hypothalamus responsible for the circadian rhythm in a mammalian brain. The SCN presumably enforces a circadian rhythm with a period of about 24 hours.

The **third ventricle** is a narrow cavity that is located between the two halves of the brain. It is one of four ventricles in the brain that communicate with one another. As with the other ventricles of the brain, it is filled with cerebrospinal fluid, which helps to protect the brain from injury and transport nutrients and waste. The circadian clock's function in this part is not clear yet [5].

The **optic chiasm** is an X-shaped structure formed by the crossing of the optic nerves in the brain [6]. The optic nerve connects the brain to the eye with over two million nerve fibers.

### 1.1 Methods

At the time the dataset was created by the researchers of Charles Allen lab, analysis of pictures was performed manually to detect the areas described above. We automate this task with unsupervised machine learning. Clustering methods divide a dataset into multiple groups to have similar data points in the same and dissimilar data points in different groups. We looked into different clustering strategies and their advantages and disadvantages [7] [8].

[1] https://www.kaggle.com/kmader/circadian-rhythm-in-the-brain

[2] https://www.ohsu.edu/xd/research/centers-institutes/oregon-institute-occupational-health-sciences/research/allen/

[3] https://www.kaggle.com/kmader/circadian-rhythm-in-the-brain/kernels

**K-Means** is one of the most basic clustering algorithms. The means of a given number of clusters are calculated and data points are assigned according to the closest mean. The algorithm is fast (runtime in $O(N)$) and only takes additional memory for the means. But the number of clusters $k$ needs to be set manually and outliers have a strong influence.

The **DBSCAN** algorithm is density-based and concentrates areas of high density into clusters. Therefore this algorithm is able to identify any shape of clusters. Unfortunately, DBSCAN calculates a full pairwise similarity matrix of the data points, which makes this algorithm very memory inefficient.

**MeanShift clustering** aims to discover dense areas in the data. It is a centroid based algorithm, which works by updating candidates for centroids to be the mean of the points within a given region. The algorithm is not highly scalable, as it requires multiple nearest neighbor searches during the execution of the algorithm. The algorithm is guaranteed to converge, however, it will stop iterating when the change in centroids is small.

**Hierarchical clustering** is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

Most clustering algorithms expect a 2D array as input. As our dataset consists of 3 dimensions, we flattened the pictures to a single dimension array. Afterwards the preprocessed dataset consists of 1685 x 65536 data points. Even with the memory resources of Kaggle Kernel this caused huge memory errors for Hierarchical Clustering and MeanShift Clustering. Additionally, even after extensive hyperparameter optimization DBSCAN failed to yield any clusters. We focused on K-Means instead as this algorithm managed to deal with the data well, especially with smaller numbers of clusters.

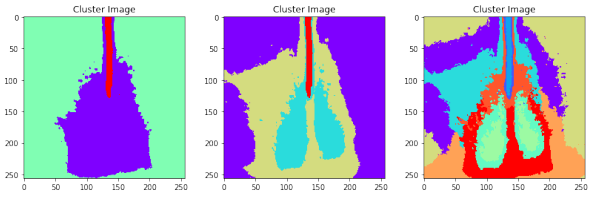



**Figure 3.** On the left side K-Means with $K = 3$, in the middle with $K = 4$ and on the right with $K = 10$

We looked into the parameter $K$ for K-Means. As we have three main clusters (the third ventricle, the SCN and surroundings) $K = 3$ seemed intuitive, but the area identified as the SCN is quite big, overlaps with the ventricle and does not differentiate the dark part in the middle (see Figure 2). With $K = 4$ the third ventricle is clearly delimited from the other parts and an additional cluster, which we call "outer SCN", developed between inner SCN and surroundings. This differentiated the very bright oval-shaped clusters from the less bright area. With rising $K$ we did not manage to capture the optic chiasm on the bottom of the picture. The fluorescence intensity of the optic chiasm does not seem to differ enough from the surroundings intensity. As the authors of the dataset hinted that the optic chiasm is in the bottom part we tried to just retrieve the lower part up to inner SCN cluster.

To see whether these clusters behave differently we selected only the pixels of these clusters and looked at the accumulated signal over time.

In Figure 4 we compare the fluorescence intensity of each cluster over time. First we filter the pixels of each cluster and take the mean as well as the first and third quartile over the course of all pictures to get the intensity signal over all seven days.

The ventricle has its very own behaviour. The clusters obtained with both $K$'s are very similar. In the first 12 hours the intensity goes down, for the next 4 days there is a strong upward trend without any noticable periodicity. For the last 3 days there is a downward trend again with a few valleys and peaks that might indicate a periodicity with a small amplitude. The authors of the dataset stated that the function of circadian clock cells in the ventricle is not clear, and that the ventricle's intensity trend may be influenced by some unrelated process.

In contrast to the ventricle the SCN shows almost no linear trend in the signal, especially in the inner SCN cluster of $K = 4$ and the cluster of $K = 3$. The outer SCN of $K = 4$ shows a moderate trend, but that might be explainable by the proximity to the surrounding cluster that has a very similar shape. There is a clearly visible periodicity that is described in detail in Section 2. The surrounding cluster has a clear upward trend. It is unclear whether there is a clear periodicity.

For the optic chiasm we selected the bottom rows of the picture up to the first row containing inner SCN cluster points. That represents the "dark band on the bottom of the picture" the scientists mentioned in their dataset description. The behaviour looks very similar to the SCN clusters, there is no clear trend and the same periodicity seems to apply. As this cluster doesn't seem to provide big insights we omitted it in further analysis.

The variance depicted by the first and third quartile is quite small throughout the clusters. The SCN cluster of $K = 3$ has a larger variance that is resolved by splitting the cluster into the clusters introduced by $K = 4$ to split the part with a clear trend from the part without a trend. We can therefore say that the clustering was successful as similar behaviours were summarized in the same cluster.

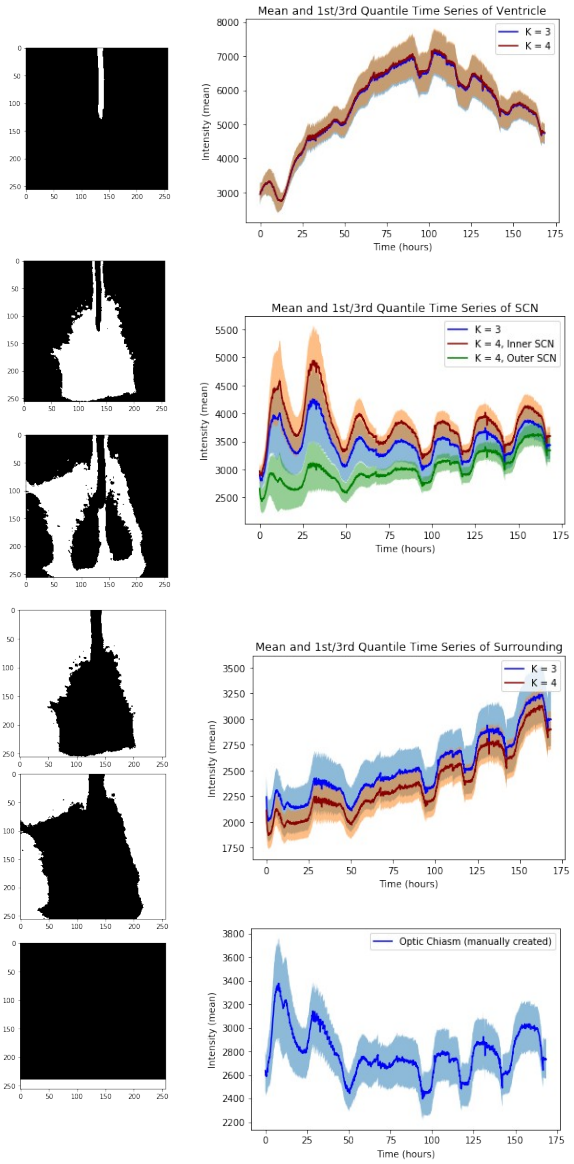



**Figure 4.** The first row shows the ventricle cluster of $K = 4$ on the left and the mean and 1st/3rd quartile of the cluster pixels signal of $K = 3$ (blue) and $K = 4$ (red) on the right. The second row shows the outer SCN of $K = 4$ on the left and the mean and 1st/3rd quartile of the cluster pixels signal of $K = 3$ (blue) and $K = 4$ for the inner SCN (red) and outer SCN (green) on the right. The third row shows the surroundings of $K = 4$ and $K = 3$ on the left and the mean and 1st/3rd quartile of the cluster pixels signal of $K = 3$ (blue) and $K = 4$ (red) on the right. The last row shows the manually extracted optic chiasm in the cluster on the left and in signal on the right.

## 2. Regularity of the Signal

The rhythm in a living organism, the inner clock, is marked by the period length, which is approximately 24 hours. This becomes clear by looking at the term ”circadian”, which means ”about a day” [9].

In different organisms and cells this rhythm can slightly vary, but the unique properties are still similar. The 24-hour-period is not dependent on external stimuli from the environment, but they can support to synchronize the period in her exactly 24 hours. The most important external influence is light [10].

The period length itself can be manipulated through drugs and hormones [11], but also the age of a living being can be a significant feature for change. Accordingly, the length of the period in humans decreases with higher age, in mice it increases. The internal rhythms are that important, that a disturbance of them can lead to mental illness and stress [12].

### 2.1 General period

For a start, it is useful to determine the overall period to have an overview of the general approximation of the 24 hours. This includes the entire intensity-averaged image with all belonging clusters.

For determining the general period, a fast and most common method is used: the **Fast Fourier Transformation** (FFT). With it, a digital signal can be broken down into its frequency components for further analysis [13]. Instead of the mean values, in this case the norm values are used with the intention to exclude the linear trend of the ventricle. Therefore the linear trend is subtracted from the mean intensity values for all given images. The functions for the period calculation by FFT were given in the kernel ”Seeing the light v2” of the Kaggle dataset [14]. They combined the FFT method with autocorrelation. The kernel result for the calculated dominant period is 24.3 hours. But it also provides an extended result, where the output is an entire array of periodic values from smaller squares in the averaged intensity image. The values in the array vary between 22.4 and 24.8 hours, which shows an approximation of the expected 24 hours.

To verify the existing kernel results, a calculation of the **standard deviations** of the period array quantifies the amount of variation in the data. So, basically it defines the spreading of the data values. The formula includes the simple average of all values (mean), which gets subtracted. This difference gets squared and the mean of those is calculated. To finish the calculation a square root of the entire formula has to be taken:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Overall standard deviation is 0.499, which can be considered a low result. A low result means, that all period values tend to be close to the **mean** value and do not vary a lot, which

indeed leads to a periodic result. The overall mean value (average) is 24.2 hours for the general period.

Another way of determining the overall period is to have a look at the **power spectrum** of FFT. There are predefined Numpy [4] functions for Python to calculate the FFT and the frequencies associated with FFT components. Basically it is given by the frequency value, which here, is transformed into period in hours ($1/Frequency$) with the highest peak on the Y-axis (*Power*), what can be clearly seen in Figure 5. The graphical result is marked by the dashed green line, which seems to be 24 hours. By calculating the corresponding period value on the X-axis numerically, the result is 23.7 hours, which is slightly shorter than the expected 24h period.

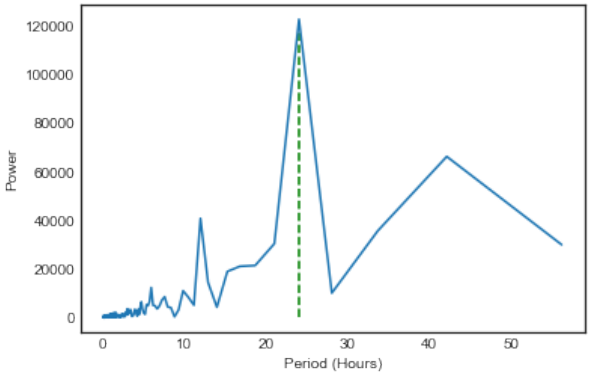



**Figure 5.** Periodicity with Power Spectrum

## 2.2 Period in different clusters

In order to understand how the general period is formed and where it derives from, the periods of the four generated clusters from the previous section are analyzed. This gives a more detailed representation about which parts of the image are important for the periodicity and which are not.

In the same way, the separate cluster periods can also be quantified with their pixels by **standard deviation** and **mean**. Each cluster has its own temporal behaviour, but it tends to be close to the expected 24 hours. As Figure 6 demonstrates, the mean period, which is closest to optimal can be seen for the outer and the inner SCN with 24.4 and 23.6 hours. The ventricle cluster has a period which is furthest away from 24 hours.

It is interesting that although the surroundings' mean period is not the expected 24 hours, the standard deviation is smaller than for other clusters. This means that the values do not vary a lot, but are mostly above the 24 hours. The standard deviation of the ventricle at 10.8 is the one with the widest spread of period values. As can be seen from the array, there are also some values which are far away from the expected result.

```
CLUSTER FOR  VENTRICLE
Periods per Pixel:
 [ 23.1  23.5  24.1 ...,  68.7  23.9  23.7]
Mean Period: 26.2839464883
Standard Deviation: 10.7835108769

CLUSTER FOR  OUTER SCN
Periods per Pixel:
 [ 23.6  23.7  23.  ...,  24.4  24.6  24.4]
Mean Period: 24.363919683
Standard Deviation: 2.21662570651

CLUSTER FOR  INNER SCN
Periods per Pixel:
 [ 23.   23.7  23.4 ...,  23.5  23.6  23.9]
Mean Period: 23.630423061
Standard Deviation: 3.42209580834

CLUSTER FOR  SURROUNDING
Periods per Pixel:
 [ 24.6  24.3  24.  ...,  24.1  24.2  24.8]
Mean Period: 24.5350736777
Standard Deviation: 0.898702953486
```

**Figure 6.** Periodicity Calculations for each Cluster

To examine which specific parts are mostly responsible for the periodicity, the fluorescence intensity over time for the four clusters is visualized in two plots. Because of the assumption that the period should be about 24 hours, the green vertical lines in this case represent an entire week divided into 24 hour sections to recognize a period.

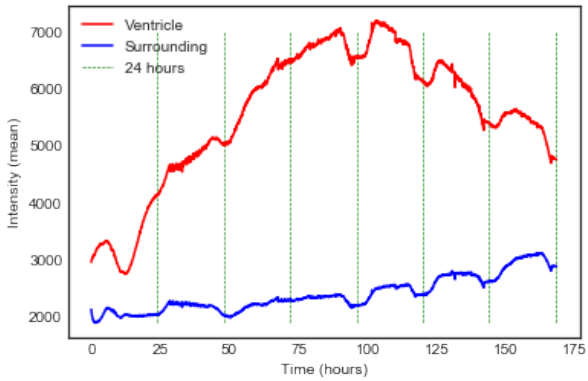



**Figure 7.** Fluorescence Intensity of Ventricle and Surrounding

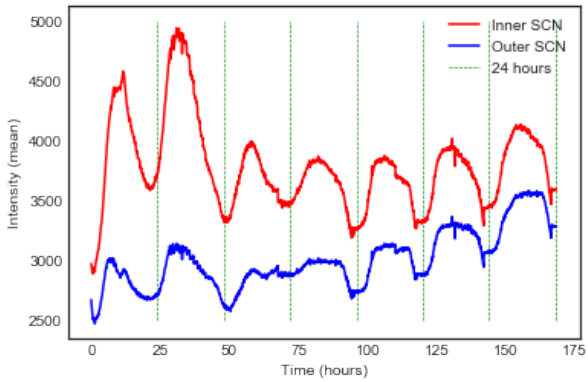



**Figure 8.** Fluorescence Intensity of Inner and Outer SCN

In Figure 7 the ventricle shows an upward linear trend over the first four days and a downward one over the last three, which was already recognized at the beginning. It has a somewhat recognizable period, but not in all 24 hour sections. It has been theorized by the authors of the dataset that some unrelated process may influence that, possibly an ongoing chemical reaction involving the colorant. In comparison to the ventricle, the surrounding region looks about as periodic, but has less of a rising trend. It stays almost flat in intensity during the entire week. It can be seen that in the last three days of the week the rhythm differs slightly from that in the first four days of the week. It can be said that those two clusters are not responsible for the main periodicity. As was already mentioned, different regions behave differently. Looking at Figure 7, there is no obvious periodicity to see for those two parts. By way of contrast, the inner and the outer SCN contains intensity changes with noticeable periods, which can be seen in Figure 8. Therefore it can be assumed, that the entire SCN is responsible for the regularity. Normally, both SCN clusters can be seen as only one. All seven peaks of both SCN functions are definitely located within the 24 hours. Therefore, in further analyses of signal regularity only the SCN part is involved, to regard specifically this periodicity in more detail.

Further, the focus of the analysis is on the **periodicity shifts** during the experiment. In the subplots in Figure 9 a pattern in the seven peaks can be recognized. In each of the seven days a peak is located. Every second peak happens at the same time (the X-axis) in the inner and outer SCN regions, but the peaks between those happen at different points in time. And the difference in intensity (the Y-axis) decreases to the end of the week, starting on the second day (24-48 hours). The highest intensity peak for the inner SCN is observed on the second day.

This can also be justified by numeric calculations, focusing only on the SCN clusters. Table 1 shows the X- and Table 2 the Y-values of the seven peaks according to both SCN functions. As already seen in the plots, it can now also be seen in the table, that the difference in time has a specific pattern. There is no difference in time of the peaks on every second day. It is not obvious whether this is simply a coincidence or a pattern, since we have only seven days of data, but it is notable that the inner and outer SCN seem to synchronize perfectly on every second day (but not exactly every 48 hours). The intensity obviously decreases.

Both parts of the SCN are somehow synchronized. The graphs peak at the same time for half of the week. There is definitely a recognizable time and phase shifting (difference). Peaks have shifting patterns in time and also in intensity.

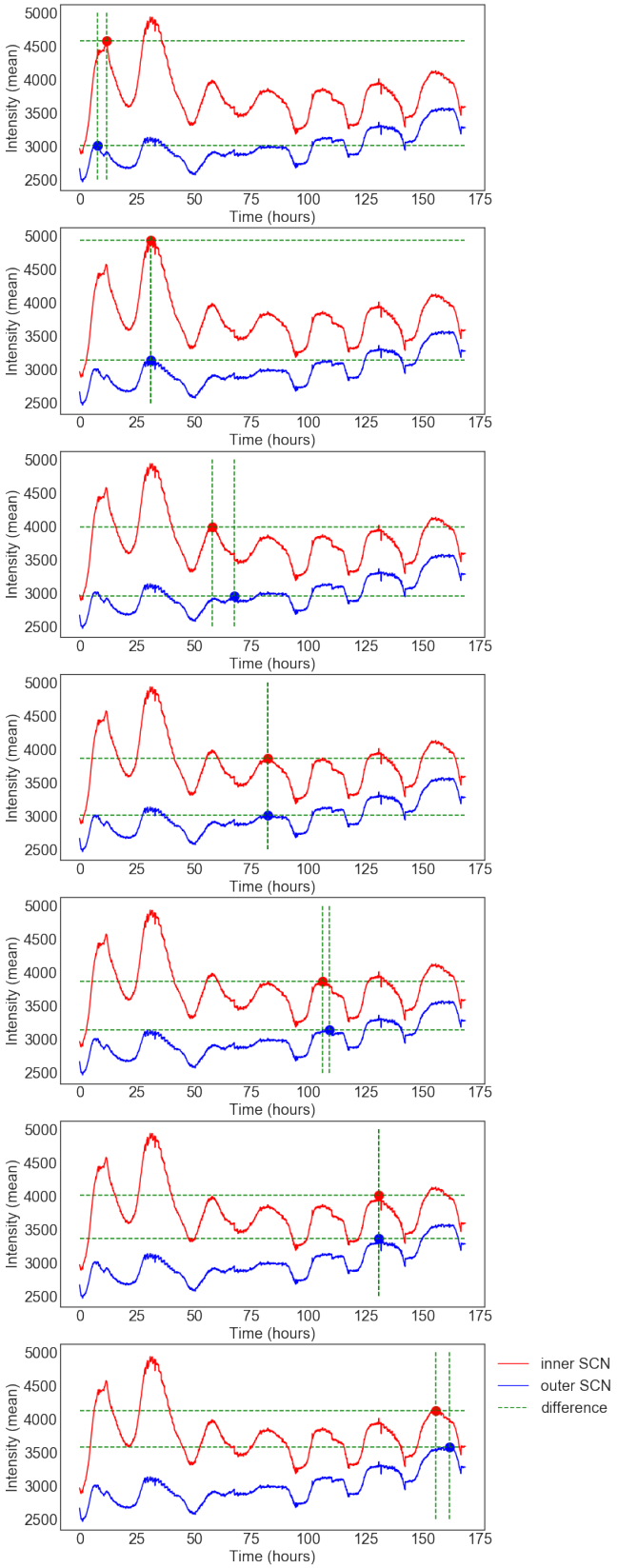



**Figure 9.** Peak Differences of Periodic Signals in Time and Intensity

**Table 1.** Numeric Peak Differences of SCN Period in Time (Hours)

| Day | Peak inner X | Peak outer X | Difference in Time (X) |
|---|---|---|---|
| 1 | 11.6 | 6.5 | 5.1 |
| 2 | 30.9 | 30.9 | 0.0 |
| 3 | 57.8 | 67.5 | 9.7 |
| 4 | 82.2 | 82.2 | 0.0 |
| 5 | 106.1 | 109.1 | 3.0 |
| 6 | 130.7 | 130.7 | 0.0 |
| 7 | 155.5 | 161.4 | 5.9 |

**Table 2.** Numeric Peak Differences of SCN Period in Intensity (Mean)

| Day | Peak inner Y | Peak outer Y | Difference in Intensity (Y) |
|---|---|---|---|
| 1 | 4577 | 3014 | 1563 |
| 2 | 4937 | 3138 | 1799 |
| 3 | 3993 | 2952 | 1041 |
| 4 | 3868 | 3017 | 851 |
| 5 | 3872 | 3139 | 733 |
| 6 | 4014 | 3363 | 651 |
| 7 | 4133 | 3577 | 556 |

## 3. Synchronization of Cells and Cell Groups

We have already established that different regions of the brain observed in the images behave differently in terms of changes in intensity, and that cells can be divided into interpretable clusters based on their intensity patterns (see Sections 1 and 2). This shows that nearby cells do form groups that oscillate together; however, there is also a question of how distance between cells, and not only their belonging to a certain cluster, influences the similarity of their behaviours.

### 3.1 Methods

We attempted to study this question in the following manner. Since we do not have the exact locations of single cells annotated, it was decided to use a simple substitute. Each image was split into non-overlapping squares of $16 \times 16$ pixels, yielding 1024 squares per image, each comprised of 256 pixels (see examples in Figure 10). We attempted to represent the expression of the gene Period1 in the cells by the average intensity of the pixels of a square at a certain time point.

As the similarity measure for a pair of squares, we use root mean squared error between the arrays of their mean intensities over the duration of the experiment. The more similar the behaviour, the smaller RMSE will be.

We treated the 256-pixel squares as points on a $32 \times 32$ grid, with distance 1 between neighboring points in the same row or column. Euclidean distance was used as the distance metric. The smallest possible distance between two squares in an image is then 1, and the largest approximately 43.84. For each range of distances, starting from $[1,2)$ and up to $[43,44)$, we determine all pairs of squares the distances between which fall into that range, and calculate the mean RMSE of all those pairs.

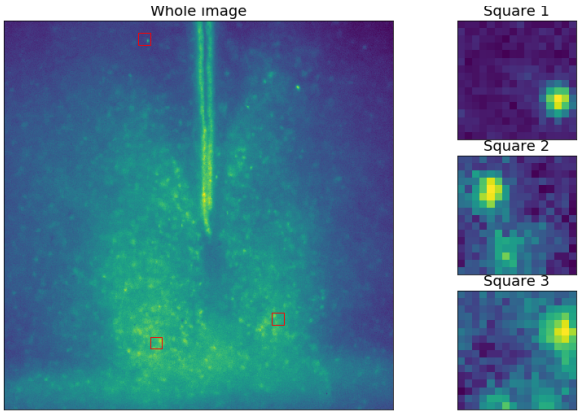



**Figure 10.** Examples of $16 \times 16$ pixel squares obtained from an image

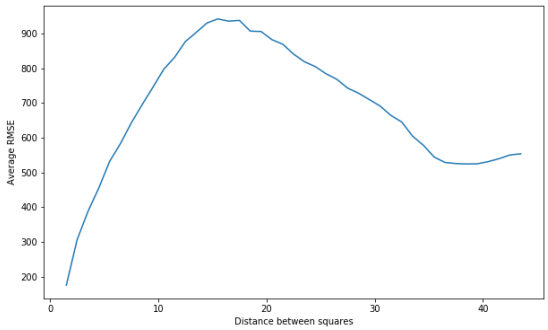



**Figure 11.** Average RMSE of all square pairs with distance within a given range

### 3.2 Results

The obtained measurements are shown in Figure 11, demonstrating how distance between cells affects the similarity of their behaviour. We can observe that with smaller distances, the larger the distance between cells, the larger the mean RMSE. However, a fall happens at distances from 16 to 38, and between 39 and 44 the average difference grows again and stagnates. We believe that this can be accounted for by the different clusters of similarly behaving cells. For square pairs within a small distance, there is a high probability that both cells belong to the same cluster, and with larger distances it becomes more likely that they are in two different clusters. However, all pairs with the largest distances will inevitably have both their squares on the edges of the image, and thus in the outer cluster (see example in Figure 12). Taking this fact into account, we can assume that the plot in Figure 11 is evidence of clusters being the most prominent factor affecting similar cell behaviour patterns.

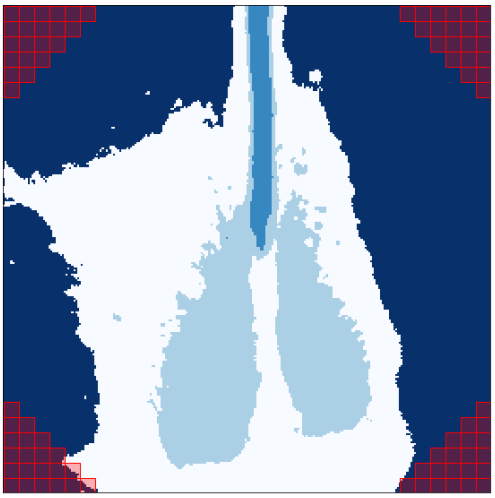


**Figure 12.** Almost all squares that appear in pairs with distance equal to or greater than 40 (marked red) belong to the outer cluster

However, does the distance within a single region matter? To answer this question, we perform the same calculations for each of four clusters (the inner SCN, the outer SCN, the ventricle and the surrounding). A square is considered as belonging to a cluster if half or more of its pixels are in that cluster, and only pairs where both cells are in the same cluster are taken into account. The result is in Figure 13.

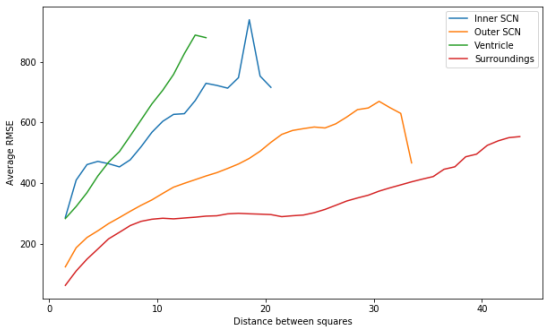



**Figure 13.** Average RMSE of all square pairs in clusters with distance within a given range

We can see that if we separate the clusters, the difference in intensity patterns still changes with distance: the closer two cells are, the more similar their behaviour. However, for all clusters except the surrounding one the similarity also decreases when the distance is at its largest. We are not sure how that fall could be explained. From the steepness of the curves we see also that the effect of distance is strongest in the ventricle, less strong in the inner SCN and outer SCN, and weakest in the outer region.

From these data we would assume that while being in one cluster will make the cells behave similarly, the closeness of cells inside one cluster will also make their patterns more alike.

## 4. Circadian Rhythm as a Decoding Task

While in previous sections we have been mostly treating the circadian rhythm and its manifestation in the brain as an encoding task, here we will look at it from a decoding point of view. We attempt to predict the time of day based on the image of the brain captured at that time. According to our data from Section 2, the period is, if not exactly 24 hours, very close to 24, which means theoretically this task should be solvable.

### 4.1 Methods

To build a time prediction model, we used random forest regression [15], as implemented in the `scikit-learn` Python library [16]. The features used are the means of intensity values of the 256 pixels of each of the $16 \times 16$-pixel squares obtained in Section 3 (see Figure 10 for examples), 1024 features in total. The value we aim to predict is $t \ mod \ 24$, where $t$ is time elapsed since the start of the experiment, in hours.

To train and evaluate the models, the 1685 instances (images) were randomly split into training and test sets (80% and 20% of the instances, respectively). Model hyperparameters were tested out using 5-fold cross-validation on the training set.

It is not optimal to use root mean squared error as the loss function in this task. The labels the model is predicting range from 0 to 24, however, since the value in the question is time of day and it is cyclical, it makes no sense to have the difference between any two values be greater than 12 hours. For instance, if the true value is 1 h, and the predicted value is 23 h, the error for the data point will be 22 h, when in terms of time the result is only 2 hours off. Because of this, we slightly modified the RMSE formula to better suit our purposes. For each instance, the error $e_i$ was calculated as either $y_i - \hat{y}_i$ or $y_i - \hat{y}_i - 24$, whichever of these has smaller absolute value, where $y_i$ is the predicted value, and $\hat{y}_i$ the actual value. The resulting loss function, which we will call $RMSE_{time}$, then is

$$\sqrt{1/n \sum_{i=1}^{n} {e_i}^2}$$

Unfortunately, it is not trivial to train `scikit-learn` models optimizing for a custom loss function. It was decided to train models using the standard MSE function implemented in the library, but evaluate performance with the custom function.

### 4.2 Results

The best model was obtained with 100 estimators, each of maximum depth 15. It showed $RMSE_{time}$ 0.780 hours on

the training set, and 1.064 hours on the test set. We do not consider these results very good, but they are satisfactory.

We inspected the importance of features in the final model, hoping to confirm our expectation that the SCN, and especially the inner SCN, should be the main contributor to the circadian periodicity. Figure 14 shows that 4 features with the highest importance scores are, indeed, located in the SCN cluster, 3 of them in its inner region.

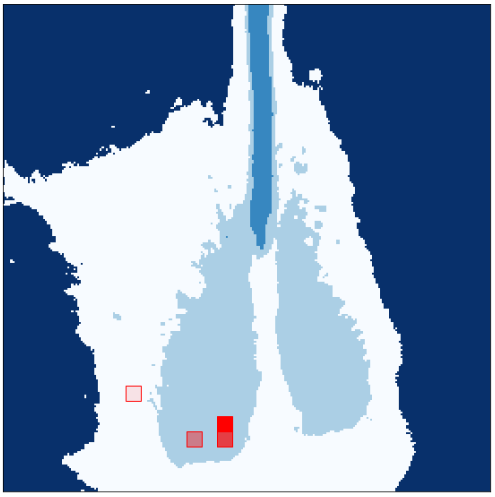
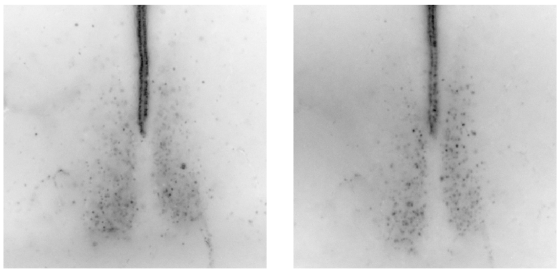


**Figure 14.** Locations of 4 most important features of the best random forest regression model (marked red, opacity decreases with increasing rank)

## 5. Cell Detection and Movement

Before tackling this question with advanced techniques and methods we started data exploration. At first, we may just look at two images from the given dataset at two different points in time (Figure 15). Here we inverted the intensity of images, so it will be easier to visually distinguish separate neuron cells. The time gap between these frames is exactly 5 days. We took this time period, because there is a dependence of the oscillations on the time of the day. The brain image differs a lot within a day and some cells can be seen at a certain time and can not be distinguished during the rest of the day (the previous sections examined oscillations and periods more deeply). As we can see in Figure 15, even though the intensity of particular neurons differs in these two frames, we still may conclude that shape of the clusters changed. This observation encouraged us for further analysis.

Some papers state that **brain cells can move** in human brain [17] and even travel for long distances in case of injury. Neurons can migrate for a long distance as well [18]. Based on these arguments we do not deny that neurons in the laboratory mouse can move.



**(a)** 43 h. from the beginning of the experiment. **(b)** 163 h. from the beginning of the experiment.

**Figure 15.** Comparison of SCN shapes in the beginning and at the the end of experiment

### 5.1 Methods

We tried to automate recognition of separate neurons and track their position over time. As we had unlabeled data (we do not know the number of neurons or their position), the task of distinguishing single neurons was quite hard. The low contrast between cell bodies and other tissues of the brain makes it even harder.

At first, let us look at the particular example of single cell movement (Figure 16). The image shows the area of the left cluster of SCN. Depending on the time of the day, the overall intensity of the image differs (that was already reviewed in the previous sections). In order to track cells movement more easily, we decided to pick images that correspond to the same time every day. The corresponding values of time after the beginning of the experiment are 43, 67, 91, 115, 139 and 163 hours. The difference between points is 24 hours. We can observe the highlighted neuron moving compared to static dashed lines. It turned out that some cells do actually move during the experiments.

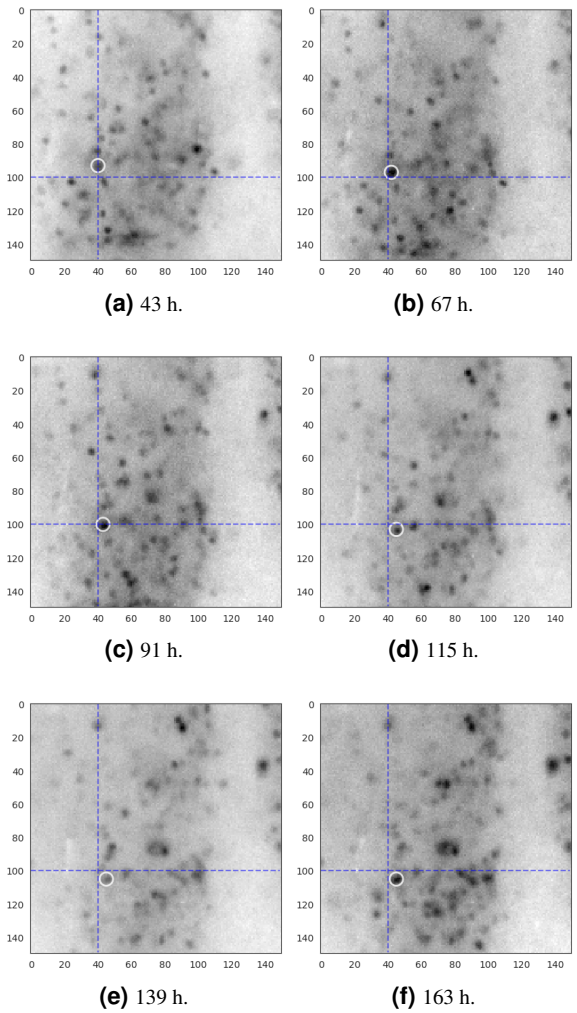
The task of distinguishing separate cells in the given dataset has the following form. Generally, neurons appear to be circular bright spots on the relatively dark background of other brain tissue. They also may form clusters – groups of neurons that are located closely together.

The most common method for recognizing the type of objects described above is **blob detection**. There are several versions of it: **The Laplacian of Gaussian** (LOG) [19], the difference of Gaussian (DOG), the determinant of the Hessian (DOH). LOG showed the best results (Figure 17).

The LOG method has several parameters to be adjusted. As we have unlabeled data we had to tune the parameters manually. Parameter values influenced the results (see Figure 17). Even though we did not succeed in distinguishing all cells with LOG, we recognized the most bright and big ones, and that should be enough to track their movement during the experiment and uncover average patterns.

The output of blob search with LOG algorithm is position and radius of the blobs. These blobs correspond to the neuron cells in our case.

We took sequence of frames (Figure 16) to build move-



**(a)** 43 h. **(b)** 67 h.

**(c)** 91 h. **(d)** 115 h.

**(e)** 139 h. **(f)** 163 h.

**Figure 16.** Neuron location at different time points during the experiment. The dashed lines are static, the neuron shifts (moves)

ment trajectory for the cells. As input image has quite low quality we did preprocessing of the images. We applied the Guassian filter [20] to smooth the images so that blob detection would be more accurate.

Then we run blob detection for every selected frame of the sequence. As a result we obtained the coordinates and radiuses of the distinguished cells. The next step was to find movement vectors of the neurons if they move at all.

To obtain the trajectory angle we performed matching blobs in one frame with detected blobs from the following frame. That gave us a starting and ending position for every time interval. We run **nearest neighbour search** based on the distance between two points (the beginning and the end of the interval) to find the next position for the detected cells. Due to image quality, the number of cells that LOG outputs differed from frame to frame. Some cells were not recognized on the following frame, so we made sure that cells will not be matched with the wrong cell by filtering out larger distances.

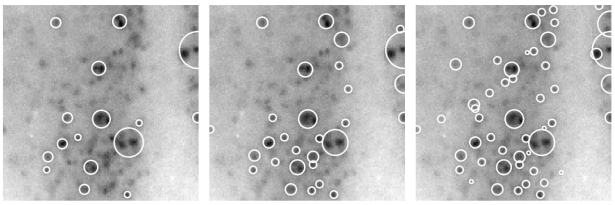


**Figure 17.** Influence of the parameters on the LOG results. Output of algorithms visualized through drown circles around distinguished cells

### 5.2 Results

After obtaining movement vectors (trajectories) we calculated the distances of these shifts (Figure 18). The values correspond to the average length of the cell movement that happened during the time interval. As we do not exactly know the real size of the observed brain area, we cannot provide real physical lengths, but we provided measurements where distance is defined in pixels. We can see that the average travel distance of the cells tends to decline toward the end of the experiment. Moreover, in the last interval, the average length is only 1.5 pixels over 24 hours.

At this stage, we can see that cells do indeed move. However, to give a more specific answer we calculated the direction of their movements (vector angle) for every interval in the sequence. Polar direction histograms are shown in Figure 19. We can see that that calculated directions are similar to the direction of the neuron in Figure 16. What is more, the number of neurons being shifted goes down with the duration of the experiment.

On the first and second polar histograms, there is a high frequency of movement in directions on average -60 degrees (or 300 degrees). This means that most of the movement that we detected was made in the same direction. There distribution of directions in the last two diagrams is more even. Considering the length of the shifts and the distribution of the direction, we can say that if there were any movements of cells, they were relatively small and mostly random or LOG blob detection imprecision caused small variation in determining neuron locations.

## Conclusion

In the present project, several directions of investigating the Circadian Rhythm in the Brain dataset have been pursued.

Several clusters comprised of similarly behaving cells have been discovered in the images of the suprachiasmatic nucleus. These clusters can be meaningfully interpreted as the inner and outer regions of the SCN, the third ventricle and the surrounding cell groups.

It was established that the general period of intensity signal is, as expected, close to 24 hours, although it can slightly vary depending on the used method. Also predictably, the central areas of the SCN were found to have the most pronounced daily periodic changes. However, the third ventricle

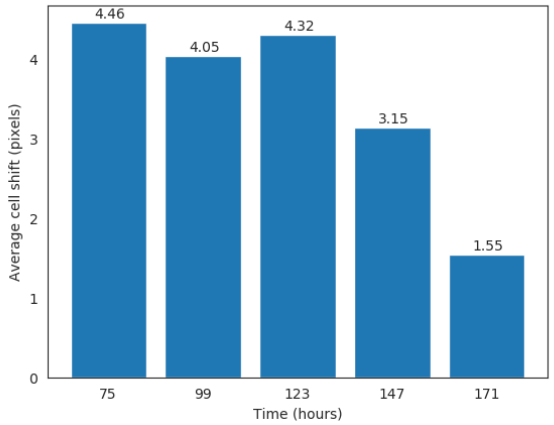
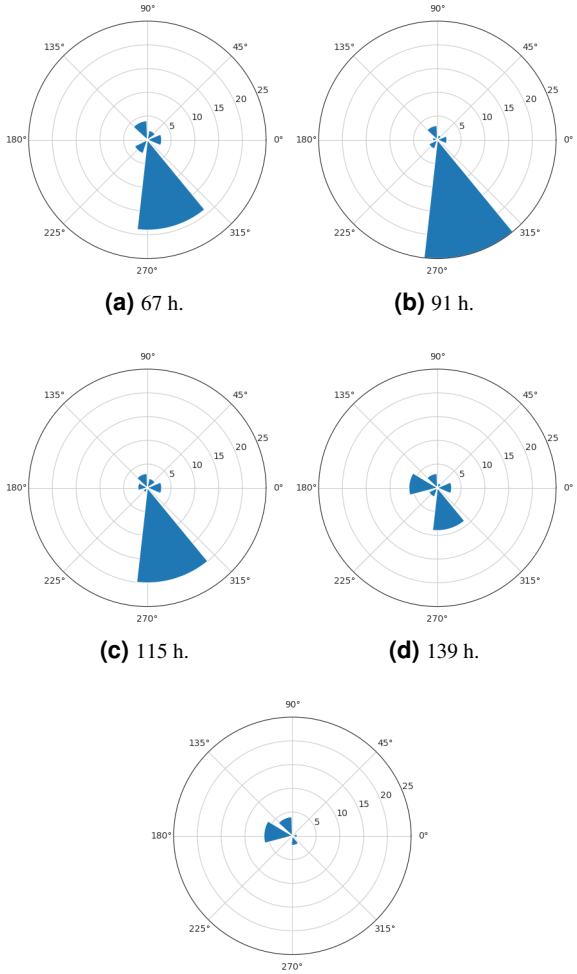



**Figure 18.** Average shift length through time intervals.

also captured in the images seems to follow its own separate trend more than it does a circadian one. Shifting patterns for the SCN in time were recognized, as well. In addition to cluster affiliation, smaller distance between cells within a single cluster was also found to cause similarity in cell behaviour patterns.

The decoding task (predicting the time of day when a picture was taken using features extracted from that picture for training a random forest model) showed satisfactory results, with best model having $RMSE_{time}$ of 1.06 hours (the *RMSE* loss function was slightly modified to better suit the task at hand).

Finally, blob detection methods were used to automatically discover individual cell locations in the images. The detected cells moved only slightly during the experiment, mostly in its beginning.



**(a)** 67 h. **(b)** 91 h. **(c)** 115 h. **(d)** 139 h. **(e)** 163 h.

**Figure 19.** Frequencies of shift directions through time intervals.

## Author Contributions

Maxi Fischer mostly investigated clustering, Katharina Fogelberg regularity and periodicity of the signal, Elizaveta Korotkova cell synchronization and the decoding task, and Kateryna Peikova cell movement. However, all authors collaborated and provided meaningful input to each other throughout the project.

## References

[1] Rachel S Edgar, Edward W Green, Yuwei Zhao, Gerben van Ooijen, Maria Olmedo, Ximing Qin, Yao Xu, Min Pan, Utham K Valekunja, Kevin A Feeney, et al. Peroxiredoxins are conserved markers of circadian rhythms. *Nature*, 485(7399):459, 2012.

[2] National Institute of General Medical Sciences. Circadian rhythms, 2017.

[3] Urs Albrecht and Gregor Eichele. The mammalian circadian clock. *Current Opinion in Genetics Development*, 13(3):271 – 277, 2003.

[4] Travis Oliphant. *Guide to NumPy*. Trelgol Publishing, 2006.

[5] L. Sakka, G. Coll, and J. Chazal. Anatomy and physiology of cerebrospinal fluid. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, 128(6):309 – 316, 2011.

[6] Desmond Kidd. The optic chiasm. *Clinical Anatomy*, 27(8):1149–1158.

[7] George Seif. The 5 clustering algorithms data scientists need to know, 2018.

[8] ScikitLearn. Documentation - clustering, 2019.

[9] Md Sahab Uddin and Abdullah Al Mamun. Circadian rhythms: Biological clock of living organisms. *Biology and Medicine*, 10(1):1–2, 2018.

[10] Rütger A Wever. Use of light to treat jet lag: differential effects of normal and bright artificial light on human

circadian rhythms. *Annals of the New York Academy of Sciences*, 453(1):282–304, 1985.

[11] Jürgen Aschoff. Circadian rhythms in man. *Science*, 148(3676):1427–1432, 1965.

[12] Katharina Wulff, Silvia Gatti, Joseph G Wettstein, and Russell G Foster. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nature Reviews Neuroscience*, 11(8):589, 2010.

[13] William T Cochran, James W Cooley, David L Favin, Howard D Helms, Reginald A Kaenel, William W Lang, George C Maling, David E Nelson, Charles M Rader, and Peter D Welch. What is the fast fourier transform? *Proceedings of the IEEE*, 55(10):1664–1674, 1967.

[14] Peter Grenholm. Seeing the light v2, 2017.

[15] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, November 2011.

[17] Christian Klämbt. Modes and regulation of glial migration in vertebrates and invertebrates. *Nature Reviews Neuroscience*, 10:769, Sep 2009. Review Article.

[18] Jeremy W Fox, Edward D Lamperti, Yaman Z Ekşioğlu, Susan E Hong, Yuanyi Feng, Donna A Graham, Ingrid E Scheffer, William B Dobyns, Betsy A Hirsch, Rodney A Radtke, Samuel F Berkovic, Peter R Huttenlocher, and Christopher A Walsh. Mutations in filamin 1 prevent migration of cerebral cortical neurons in human periventricular heterotopia. *Neuron*, 21(6):1315 – 1325, 1998.

[19] Wikipedia contributors. Blob detection — Wikipedia, the free encyclopedia, 2018. [Online; accessed 21-January-2019].

[20] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Prentice Hall, Upper Saddle River, N.J., 2008.