**Chapter 1**

**1 Introduction**

Uncertainty visualization is an ongoing area of research but a topic that many people avoid due to the additional complexity that it introduces. There are various studies conducted for uncertainty representations, for example: textual representation such as captions or tooltips [51], graphical representations such as glyphs [21, 54], custom color palettes such as VSUP [35], bivariate choropleth maps [43], texture patterns [29] and so on. But as far we know, no uncertainty representation has used Chromatic Aberration. We introduce machine learning model uncertainties as chromatic aberration in visual interfaces. To accomplish the purpose, we have categorized the scope of the research with several core components: firstly, collect relevant data from some reputable sources. Secondly, generate uncertainty information from predictions based on the data (accomplished by feeding collected data into machine learning models and calculated from the resultant forecasts [6]). Thirdly, visualize the uncertainty and data using chromatic aberration, as well as competing existing methods. Fourthly, conduct a controlled human-computer interaction experiment to evaluate the effectiveness of the new visual representation. Fifthly, explain experimental results with numerical analysis and draw conclusions.

**1.1 Background and Motivation**   
The outbreak of coronavirus COVID-19 first emerged in China in December 2019 and the expansion has propagated all over the world, being declared as an international public health crisis by WHO. Since then, the world has been very affected in almost all respects. Various preventive health measures were and are imposed, and different short-term restrictions are applied to the habitants in different countries at different times. But the mortality rate was not mitigated significantly until immunizations started and, tragically, over 318 million people have been infected and 5.5 million have died the world over. The infection and death rate have oscillated in different countries due to a variety of reasons. Moreover, the strain of the virus is changing frequently in different geographical locations with more power and variations and a few of the variants like the British variant, the Delta variant, the Indian variant and most recently the Omicron became the prime concern for the world community. Though a great deal of research is being conducted and wide range of immunization processes have impacted the severity of the pandemic, still at the time of writing this thesis, nobody knows when the world will be rid of this severe pandemic and return to normal life again.

Recently, many studies have been conducted to forecast the trend of the spread of the COVID-19 pandemic using various statistical models as well as machine learning models. The autoregressive integrated moving average (ARIMA) model has been widely used in previous studies to analyze and predict the spread of the diseases such influenza [1], Cholera [5], along with many other popular machine learning algorithms [2, 3, 9]. The pandemic started very abruptly and so during the first year, it was difficult to develop efficient systems to forecast trends due to the lack of required data. But after more than one year, we have data to explore, analyze and forecast with the help of modern machine learning algorithms. The ability to identify the expansion rate at which the disease is spreading is very important to confront it and help governments’ regarding contingent policymaking to properly address the consequences of the pandemic and encourage people to be cautious and follow the rules and health guidelines to achieve the maximum benefit by saving valued lives. That’s why one of the objective’s behind the current research is develop new tools for uncertainty visualization. We use property driven predicted results of COVID-19 as a test case for exploring chromatic aberration as a visual representation of uncertainty. If we can develop more effective representations of uncertainty, then it might help community administrators with planning or at least improve the means of communication with the general public. And more generally, the development of better uncertainty visualizations could be of use in many other areas as well.

**1.2 Background** **Concepts**We will now introduce related terms used in the dissertation so that the reader can better understand the work.

**1.2.1 Machine Learning (predictive models)**

Machine learning is an approach of artificial intelligence (AI) to provide automatic learning through the uses of data. What separates this from other solutions is it does not need explicit programming to perform the task since the algorithms are designed to themselves learn from data. There are three types of machine learning algorithms i. **Supervised Learning** (In this type, the machine learning algorithm is trained on labeled data. Even though the data needs to be labeled accurately for this method to work, supervised learning is extremely powerful when used in the right circumstances) ii. **Unsupervised Learning** (This is a type of algorithm that learns patterns from untagged data. This type of learning does not have labels to work off, resulting in the creation of hidden structures. Relationships between data points are perceived by the algorithm in an abstract manner, with no input required from human beings.) iii. **Reinforcement Learning** (This learning directly takes inspiration from how human beings learn from data in their lives. It features an algorithm that improves upon itself and learns from new situations using a trial-and-error method).

We have chosen three supervised learning algorithms (MLP, CNN and LSTM). Along with supervised learning we have also chosen another statistical model (ARIMA). We discuss further detail about these algorithms in Chapter 3.

**1.2.2 Streamgraph**Stream graphs are an approach to visualization which are ideal for displaying high-volume datasets, to discover shapes, trends, and patterns over time across a wide range of numerical groups side by side. For example, seasonal peaks in the stream shape can suggest a periodic pattern. They work even better when there is an interactive component involved that enables the following of each separate “flow” or allow filtering the view in some way. The following example shows number of deaths count among the continents for the duration of 10 days.

Chart

Description automatically generated

Figure-1: Streamgraph (ref https://app.flourish.studio/visualisation/4023285)

**1.2.3 D3.js**

D3 is a JavaScript library for manipulating web documents based on data. It creates visualizations by binding the data and graphical elements to the Document Object Model and eventually produce dynamic and interactive data visualizations in web browsers with the help of standard web technologies like HTML, CSS, SVG. The visualizations developed in this thesis were all created using the D3 visualization library.

**1.2.4 Uncertainty**

Uncertainty is an essential part of life and is defined by lack of sureness or certainty in data. The lack of certainty is a state of limited knowledge where it is impossible to exactly describe the existing state or a future outcome. In practice, uncertainty is a complex concept and there are many kinds of uncertainty that decision makers must face. It covers a broad range of concepts like inconsistency, doubtfulness, reliability, inaccuracy, or error (unknown or not quantified error). Hence, it is difficult to give a generally accepted definition of uncertainty [45]. Uncertainty describes a comparison that can most clearly be understood visually, such as the difference between surfaces generated using different techniques, or a range of values that a surface might fall in. A simple approach to the visualization of this type of information is a side-by-side comparison of data sets [48]. Different types of uncertainty result in differing interpretations and misinterpretations and so different people perceive and explain it differently, for example: participants in a survey used phrases like ‘imperfect knowledge,’ ‘inadequate information’ and ‘lack of absolute knowledge’ to describe uncertainty. Some participants saw uncertainty as a time when the probability of something is not 1.0. When more than one event could happen, this was uncertainty. One participant articulated this as a ‘partial belief’ in something [53].

Data uncertainty is the degree to which it is inaccurate, imprecise, or unreliable. It can come from source (e.g.: data provider), data lineage (e.g.: from calculation), noise (e.g.: inaccurate post in social media), abnormalities (e.g.: two sources give different values) to name a few. We are considering only the uncertainties calculated from machine learning model predictions.

**1.2.5 Texture**Texture is the perceived surface quality of a work of art. It can be used in the analysis of images or charts in several ways: in the segmentation of scenes into distinct objects and regions, in the classification or recognition of surface materials, and in the computation of surface shape. It has been studied extensively in the field of computer vision, computer graphics, and modeling the low-level human visual system in cognitive psychology. Researchers have used different methods to study the perceptual features inherent in a texture pattern [22, 25, 56].

it is advantageous to

consider interdisciplinary integration of these research ef-

forts and apply it in new areas, e.g., data visualization.

In the visualization field, people have studied methods for using texture patterns to display information. Although different group of people concentrate on different tasks, it is advantageous to consider interdisciplinary integration of these research efforts and apply it in new areas, e.g., data visualization [57]. Textures can be generated in different ways but since our research work is implemented in web, we have used the JavaScript and CSS driven textures called SVG patterns. The SVG <pattern> element allows us to define patterns inside of our SVG markup and use those patterns as a fill. Each pattern has specific shape and we have mostly used circle and rectangle pattern to represent our texture. We will further discuss the generation procedure and algorithm in chapter 3.

**1.2.6 Chromatic Aberration**Chromatic aberration is a color distortion or alteration that is sometimes noticed on high contrast edges of objects in photographs. Since different colors of light refract to the different angles upon traveling through materials with refractive indices [9] (Figure 1), the resulting images may appear to be distorted [10]. It happens when the light of certain wavelengths becomes bent. It usually appears in the form of purple, red, blue, cyan, green fringes. It can be seen alongside deep contrast edges and traditionally it means finding colors where they should not be or found in an unexpected form of color.

**Chart

Description automatically generated A picture containing plant, tree

Description automatically generated   
Figure 2: Examples: Left - [10], Right - [expertphotography.com](https://expertphotography.com/remove-chromatic-aberration-photoshop/)**

In figure 2, we see two forms of CA where the left one shows how chromatic aberration occurs in optics as an effect when a lens is not able to properly refract all the wavelengths of colour in the same point. On the other hand, the circle bounded area on right picture shows how the quality of the picture subtly distorted.

CA is a phenomenon that can cause image distortions when viewed through lenses. Since light of various colors refract at various angles on traveling through materials with refractive indices (Figure 2-left), the resulting images may appear to be distorted. Since more and more people undergo impaired vision due myopia or astigmatism, the usage of corrective lenses increases, making more people vulnerable to this type of visual distortion. Rationally many displays use three colors (RGB) of light, because it provides a convenient conversion process between human color vision and the color space and hence it creates a very special phenomenon where the misperception comes from aberration of three distinct lights [10]. Conforming to the aberration formation concept, we have chosen three color (RGB) channels to form a blended shapes (circle, rectangle, etc.) where they are internally laterally shifted from each other by the amount of uncertainty.

CA is a problem of an image quality so most of the research about CA are conducted to fix the problem and improve image quality thereby. On the other hand, uncertainty is the problem of data quality and relevant research are conducted mostly regarding reducing it to improve data certainty. And some of the research are conducted to visualize uncertainty with traditional approaches such glyphs, opacity, and so on. Since our goal is neither to improve image quality nor data quality, we borrowed the term CA for our research to represent uncertainty as a novel approach in the field of visualization.

**1.3. Problem statement**   
The primary objective of this research is to present and evaluate a novel concept of employing CA to represent uncertainties. For our test case we use uncertainty values generated from predictive machine learning algorithms by amassing and feeding the COVID-19 data into the models. We hypothesized that our proposed system would potentially offer a more effective means of visualizing this type of information.

To implement the system, we needed to consider the following aspects:

1. How to generate the realistic uncertainty data?
2. Which platform or framework to be chosen to implement the visualization?
3. What is the design process of representing uncertainty with CA?
4. How to evaluate CA representation?
5. What is applicability of this representation?

Considering the above aspects, we have chosen to use recent WHO authorized COVID data to feed into three machine learning predictive models and one statistical model to obtain forecasted results for a certain period [3, 6]. Then calculated uncertainties from the predicted results and those are depicted as CA in D3 based visualizations as well as existing alternative options such blur, noise, and palette-based uncertainty visualizations [35]. We conduct a comparative user study and conduct numerical analysis to assess the effectiveness of our novel design of uncertainty representation with CA. The survey is conducted online given potential issues with in-person contact during the pandemic.

**1.4. Approach**

At the first step we sought a suitable dataset in terms of completeness and accuracy. By analyzing numerous data repositories, we determined that the WHO approved OWID dataset is the most comprehensive one among all others.

Secondly, we had to study an extensive set of existing work about forecasting from temporal data using machine learning models and chose four popular modeling algorithms for our research. Since, finding and comparing the effectives of algorithms’ is out of our scope of work, we randomly chose a reasonable set of the models because we needed to generate the uncertainty data for the countries by using the predictive models and ignoring all inherent uncertainties itself.

Thirdly, having the data generation component in python, we needed to write APIs to connect and pull the data when drawing the charts. Since the model training and data generation for all countries are long running processes, we precompiled the models to generate the data and stored the data into json files so that they can be input readily and sent back to the client on demand.

Fourthly, we have chosen D3.js as our front-end library for drawing the charts because it is an efficient platform for visualization prototyping and widely used. Since developing the basic drawing algorithms is not our goal, we relied on the existing library features but the aggregate data collection, preparation, manipulation, correction and drawing algorithms were developed specifically for this thesis.

Fifthly, we conducted an experiment to evaluate the approach approved by the Research Ethics Board (REB) of Dalhousie University and with the participation of the members of the community.

Finally, in we conduct a numerical analysis and offer a discussion on the survey responses and compare alternative perspectives of reference studies to consolidate and explore the research outcomes.

**1.5. Thesis outline**

The remainder of this thesis is organized as follows. In **chapter 2**, we review the relevant literature on Predictive Machine Learning Models, Texture, Uncertainty, and CA. The literature review is subdivided into several sub-sections based on the contents. **Chapter 3** presents data processing, introducing predictive machine learning algorithms and necessary arrangement to setup models, brief description of time series forecasting, snapshots of uncertainty data. **Chapter 4** focuses on user study and numerical analysis for the sake of evaluation. **Chapter 5** shows the example of uses of CA in different charts. Finally, in **Chapter 6**, we discussed and summarized the thesis content, mentioned limitations, and suggest potential directions of future work and associated improvement.