**ITCS451, Semester 1/2023**

**Bayesian Network Project**

**<u>Group members</u>**

| No. | Student ID | Student name | Section |
|---|---|---|---|
| 1. | 6488179 | Ponnapassorn Iamborisut | 1 |
| 2. | 6488181 | Thadeeya Duangkaew | 1 |
| 3. | 6488210 | Ravikarn Jarungjitvittawas | 1 |

1. Description of the problem you are modeling.
   The problem that we are modeling of the Bayesian network is to understand and predict the health status or disease of infants based on several key factors. These factors encompass a range of medical conditions, physiological parameters, and diagnostic indicators. The variables in the network are interconnected, reflecting potential dependencies and relationships between different aspects of infant health.

2. Did you use real data? (10% extra credit)

   ☑ Yes ☐ No

   If yes, describe your data in the following table.

| | |
|---|---|
| **Source of data** | https://www.kaggle.com/datasets/bhavkaur/synthetic-infant-health-dataset/data |
| **Type of data** | ☐ raw data<br>☑ published statistics |
| **Modification to the data (If any)** | We did not modify the data. |

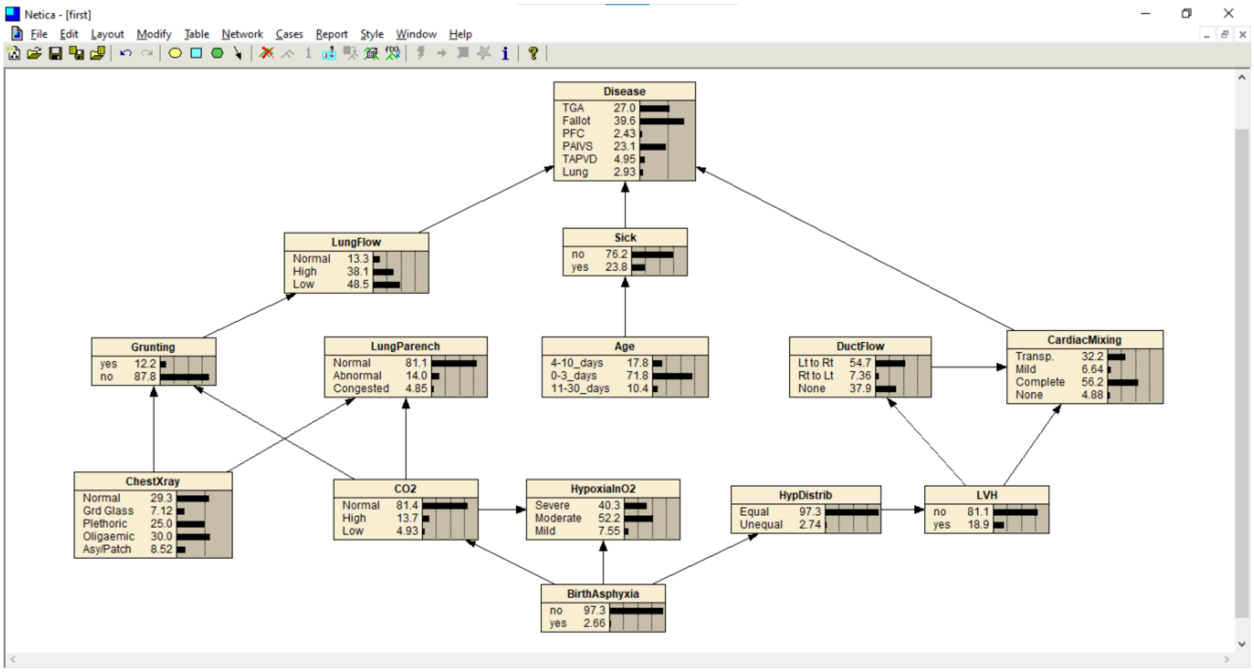
3. Bayes net description.
   a. Bayes net description.

| | |
|---|---|
| How many nodes (target and non-target)? | There are 14 nodes in the Bayesian network, of which 13 are non-target nodes and 1 are target nodes. |
| What is your target node(s)? | The node "**Disease**" is the target node. |
| What are the relationships between the target node and the other nodes and why? | <u>Direct Causal Relationships:</u><br>- **BirthAsphyxia** causes **CO2**, **HypoxiaInO2**, and **HypDistrib**.<br>- **CO2** causes **Grunting**, **LungParench**, and **HypoxiaInO2**.<br>- **ChestXray** causes **Grunting** and **LungParench**. |

| | |
|---|---|
| | - **Grunting** causes **LungFlow**.<br>- **LungFlow** causes **Disease**.<br>- **Age** causes **Sick**.<br>- **HypoxiaInO2** causes **LVH**.<br>- **LVH** causes **DuctFlow** and **CardiacMixing**.<br>- **DuctFlow** causes **CardiacMixing**.<br><br>Indirect Causal Relationships:<br>- **BirthAsphyxia** contributes to **Disease** through its influence on **CO2**, **HypoxiaIn02**, and **HypDistrib**.<br>- **CO2** further contributes to **Disease** through its effect on **Grunting** and **LungParench**.<br>- **ChestXray** indirectly influences **Disease** by affecting **Grunting** and **LungParench**.<br><br>Non-Causal Relationships:<br>- **Sick** and **Lung Parench** are both associated with **Disease**, but there is no direct causal relationship between them. They may share common underlying causes or be part of a complex causal chain.<br>- **BirthAsphyxia**, **Sick**, and **Lung Parench** all contribute to Disease, but their effects are not additive. The specific impact of each factor depends on the context and the individual's overall health condition.<br><br>This network highlights the intricate interplay between various factors that influence the development of "**Disease**". Understanding these relationships can aid in early diagnosis, risk assessment, and treatment interventions. |
| Are there any relationships among the non-target nodes? Why are these needed? If not, why not? | Yes, there are several relationships among the non-target nodes in the Bayesian network. These relationships are important for understanding the overall causal structure of the network and for making accurate inferences about the target node.<br><br>**For examples** of relationships among the non-target nodes:<br>- CO2 and HypoxiaIn02 are both measures of oxygen levels in the blood. They are therefore highly correlated, meaning that they tend to vary together. This correlation is important because it allows us to make more accurate inferences about one of these nodes if we know the value of the other node.<br><br>- ChestXray, Grunting, and LungParench are all measures of lung function. They are therefore also correlated to some extent. This correlation is important because it allows us to identify potential problems with lung function by looking at multiple different measures. |

| | |
|---|---|
| | - Age and Sick are related because infants 0-3 days are more likely to be sick. This relationship is important because it allows us to take age into account when making inferences about the likelihood of disease.<br><br>- BirthAsphyxia, Hypoxialn02, LVH, and DuctFlow are all related because they are all potential complications of birth asphyxia. This relationship is important because it allows us to identify infants who are at risk of developing these complications and to provide them with appropriate care.<br><br>In conclusion, the relationships among the non-target nodes are an important part of the Bayesian network. They help us to understand the causal structure of the network, make more accurate inferences about the target node, and identify potential confounders. |

b. Bayes net screenshot



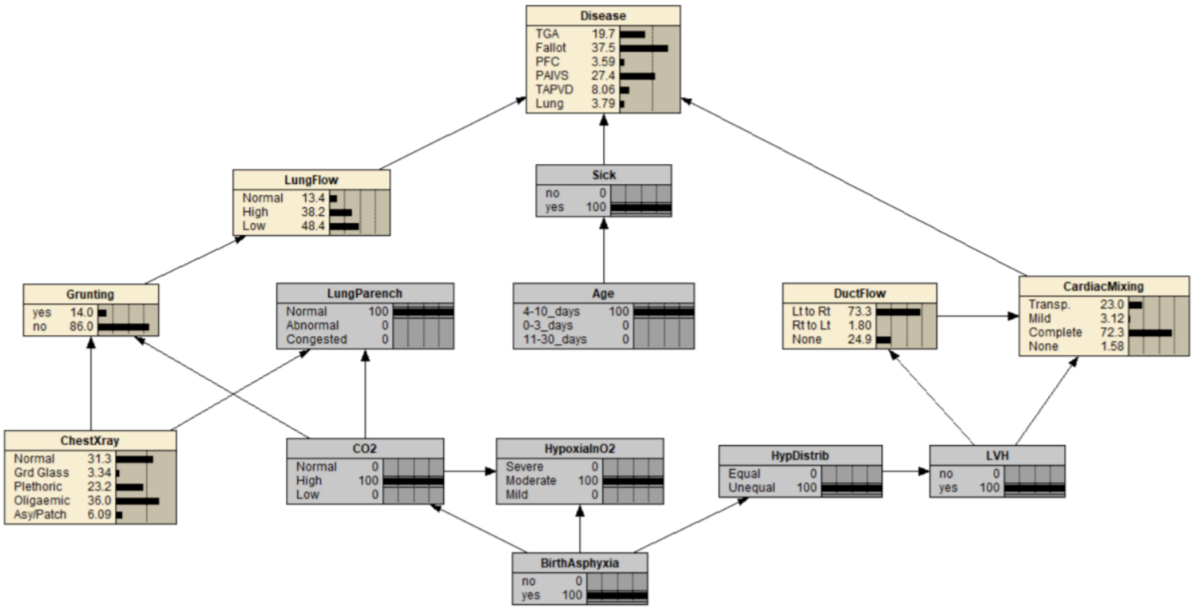

4. Explanation of your network nodes

| No. | Node Name | Node Description |
|---|---|---|
| 1. | BirthAsphyxia | This node represents the presence(yes) or absence(no) of birth asphyxia, which is a condition that occurs when an infant does not get enough oxygen during or shortly after birth. Birth asphyxia can cause damage to the brain and other organs. |
| 2. | HypDistrib | This node represents the distribution of oxygen in the infant's body. It can be either equal or unequal. Unequal distribution of oxygen can lead to hypoxia, which is a condition in which the body does not get enough oxygen. |
| 3. | HypoxiaInO2 | This node represents the presence or absence of hypoxia in the infant's oxygen levels. Hypoxia can be severe, moderate, or mild. |
| 4. | CO2 | This node represents the carbon dioxide levels in the infant's blood. Carbon dioxide is a waste product of metabolism. High levels of carbon dioxide can be a sign of respiratory problems. |

| | | |
|---|---|---|
| 5. | ChestXray | This node represents the findings on the infant's chest X-ray. The X-ray may show normal findings, ground-glass opacities, plethoric lungs, oligaemic lungs, or asymmetric/patchy lung markings. |
| 6. | Grunting | This node represents the presence or absence of grunting sounds made by the infant. Grunting is a sign of respiratory distress. |
| 7. | Disease | This node represents the presence or absence of a disease that could be causing the infant's symptoms. The diseases that are considered in the Bayesian network are:<br>- Transposition of the great arteries (TGA)<br>- Tetralogy of Fallot (Fallot)<br>- Pulmonary foramen closure (PFC)<br>- Pulmonary artery inlet valve stenosis (PAIVS)<br>- Total anomalous pulmonary venous drainage (TAPVD)<br>- Lung disease |
| 8. | Age | This node represents the age of an infant in days. Age is a risk factor for many diseases, including respiratory infections and congenital heart disease. |
| 9. | LVH | This node represents the presence or absence of left ventricular hypertrophy (LVH). LVH is a condition in which the left ventricle of the heart is thickened. LVH can be caused by a variety of conditions, including high blood pressure, heart valve disease, and cardiomyopathy. |
| 10. | DuctFlow | This node represents the flow of blood through the ductus arteriosus (DA). The DA is a blood vessel that connects the pulmonary artery and the aorta. It is normally present in infants before birth, but it closes shortly after birth. If the DA remains open, it can cause problems such as congestive heart failure. |
| 11. | CardiacMixing | This node represents the mixing of blood in the heart. It can be normal, mild, complete, or absent. |

| | | |
|---|---|---|
| 12. | LungParench | This node represents the state of the lung parenchyma, which is the tissue of the lungs. It can be normal, abnormal, or congested. |
| 13. | LungFlow | This node represents the flow of air through the lungs. It can be high, normal, or low. |
| 14. | Sick | This node represents whether or not an infant is sick. This node is the target node of the Bayesian network. It is used to predict whether or not an infant is sick, based on the values of the other nodes. |

5. Two examples of inference: Enter some evidence and show the updated probabilities of the main hypothesis node(s). Explain what each example is showing.
    a. Example1
        - Screenshot:



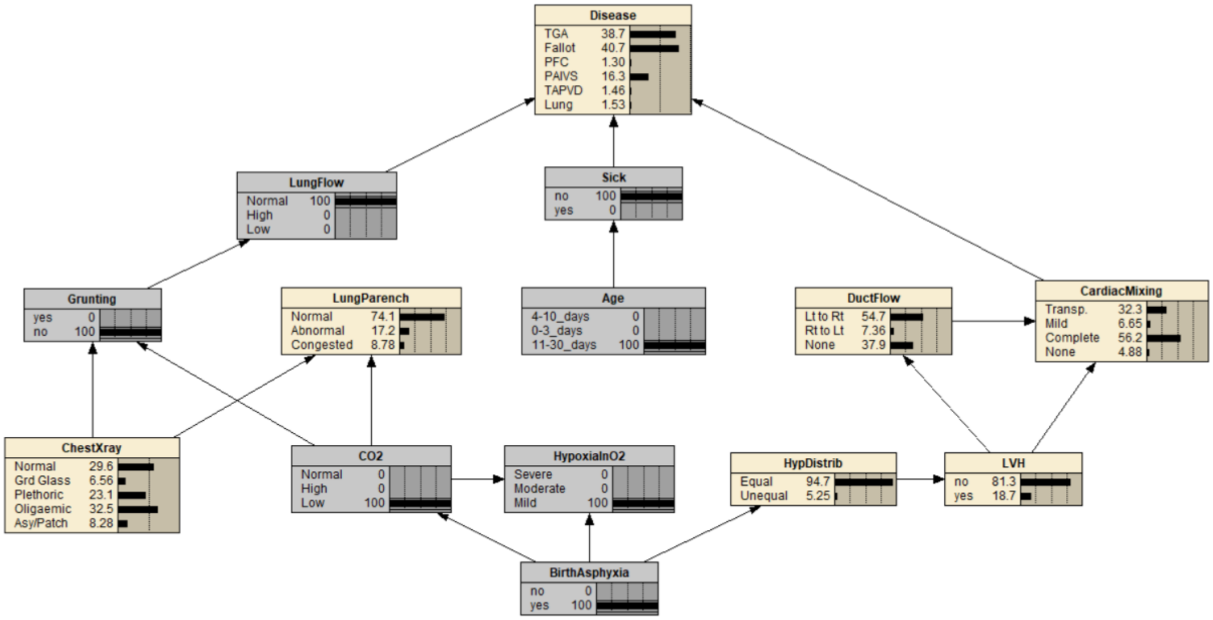

        - Explanation. What is the scenario? Why does the result make sense?

In this case these infant ages 4 - 10 days are born with Asphyxia and have high levels of carbon dioxide (CO2) in the blood, causing symptoms of Hypoxia and causing left ventricular hypertrophy (LVH). But this infant still has lung parenchyma in the normal level, which was diagnosed as a sick infant.

The bayes network predicted that this infant has a 27.0% chance of having Transposition of the great arteries (TGA), has a 37.5% chance of having Tetralogy of Fallot, has a 3.59% chance of having persistent fetal circulation (PFC), has a 27.4% chance of having Pulmonary atresia with intact ventricular septum (PAIVS), has a 8.06% chance of having Total anomalous pulmonary venous connection (TAPVC), has a 3.79% chance of having lung disease.

b. Example2

- Screenshot:




- Explanation. What is the scenario? Why does the result make sense?

In this case these infant ages 11 - 30 days are born with Asphyxia and have low levels of carbon dioxide (CO2) in the blood, causing symptoms of Hypoxia and causing grunting. But this infant still has lung flow in the normal level, which was diagnosed as a non-sick infant.

The bayes network predicted that this infant still has a 38.7% chance of having Transposition of the great arteries (TGA), has a 40.7% chance of having Tetralogy of Fallot, has a 1.30% chance of having persistent fetal circulation (PFC), has a 16.3% chance of having Pulmonary atresia with intact ventricular septum (PAIVS), has a 1.46% chance of having Total anomalous pulmonary venous connection (TAPVC), has a 1.53% chance of having lung disease.

6. Describe briefly what you could do to improve your network.

To improve our network, we could review and refine the set of features used in the network. Consider additional relevant features that might contribute to a more accurate prediction. Assess the impact of each feature on the network's predictions and remove redundant or irrelevant ones. And might split our dataset into training and testing sets to validate the performance of the Bayesian network. Use cross-validation techniques to assess the generalization capabilities of the model.